PENGARUH PANJANG TES DAN UKURAN SAMPEL TERHADAP KEKEKARAN ESTIMASI PARAMETER PADA TEORI RESPON BUTIR (*ITEM RESPONSE THEORY*)

Oleh: Sukirno DS
FISE Universitas Negeri Yogyakarta

Abstract

Developing fair and objective test had been developed in many ways, but they would solve huge educational problem. Simulation study conducted by Hambleton and Cook, inspired this study to looked for evidence the effect of test length and sample size on the robustness of item parameters estimation. This article gave more evidences in describing the effect of test length and sample size on robustness of item parameters estimation in norm reference evaluation of item response theory.

Hypothetical data used in this study was generated by DGEN application program. Test length was modeled in 15 and 55 items that represent short and long test form. The sampel size effect was tested by three different sampel sizes, 271, 605, and 1.999 to estimate 1, 2, and 3 paramters according to the rule of thumb. By BILOG program, the data was estimated and presented in table and graphic, product moment correlation, mean square error, standard error, information function curve, and univariate and multivariate analyses.

Based on data analysis, it is indicated that: 1) the test length affected robustness of discrimination and pseudo guessing parameters estimation, but had no effect on difficulty parameter estimation; 2) adding sample size decreased the mean square error and standard error discrimination and difficulty parameters estimation, but had no effect on pseudo guessing parameter

estimation; 3) test length had no effect on information function score, but sample size had effect on information function score.

Key words: test length, sampel size, parameter estimation robustness, simulation

Pendahuluan

Latar Belakang Masalah

Kalimat kritis yang ditulis dalam New Yok Times berbunyi "*the more we compete the less we fight*" sangat pantas apabila diadopsi sebagai slogan dunia tes. Kompetisi merupakan kunci sukses sejak tahun 1990 bagi dunia industri, negara, atau institusi lainnya, karena kompetisi merupakan prakondisi yang harus ada agar institusi dapat berkembang, saling memiliki rasa ketergantungan dan pemahaman terhadap Negara atau institusi lain (Olmedilla,1992). Tanpa dilandasi kompetisi akan timbul konflik politik, stagnasi ekonomi, dan kericuhan sosial, demikian pula dalam dunia pendidikan. Dalam setiap kompetisi dibutuhkan yuri yang memiliki posisi strategis. Dalam konteks pendidikan, kompetisi dimanifestasikan dalam bentuk evaluasi.

Evaluasi sebagai upaya mengukur dan menilai keberhasilan pengajaran yang dilaksanakan menduduki posisi yang tidak kalah penting dari kegiatan pengajaran itu sendiri. Berbagai keputusan pendidikan yang berupa keputusan diagnostik, bimbingan dan konseling, tes penempatan (*placement test)*, serta kelulusan, diperoleh melalui kegiatan pengukuran.

Ada dua klasifikasi besar teori pengukuran yaitu teori klasik dan *item response theory* (teori respon butir). Aplikasi teori respon butir sudah dipakai meluas, namun dalam praktiknya masih banyak ditemui kendala di lapangan, misalnya model teori respon butir tidak cocok untuk penilaian klasikal dengan jumlah sampel kecil, rumus yang digunakan rumit, harus memenuhi asumsi teori respon butir (pertama, performansi subjek pada suatu item dapat diprediksikan

oleh seperangkat faktor yang disebut *latent traits* atau kemampuan, dan kedua, hubungan antara performansi subjek pada suatu item dan perangkat kemampuan laten yang mendasarinya digambarkan oleh fungsi naik monoton dari *item characteristic curve*, kesulitan pemilihan model parameter yang cocok, apakah model 1, 2, atau 3 parameter yang akan digunakan.

Teori pengukuran klasik telah banyak berjasa dalam dunia pengukuran dan bahkan masih digunakan sampai sekarang. Namun demikian dalam teori pengukuran klasik terdapat keterbatasan karena bersifat *group dependent* dan *item dependent* (Hambleton dan Swaminathan, 1991; serta Azwar, 1999).

Group dependent artinya hasil pengukuran tergantung dari kelompok peserta yang mengerjakan tes. Jika tes diujikan kepada kelompok peserta dengan kemampuan tinggi, tingkat kesulitan butir soal akan rendah. Sebaliknya jika tes diujikan kepada kelompok peserta dengan kemampuan rendah, tingkat kesulitan butir soal akan tinggi.

Item dependent artinya hasil pengukuran tergantung dari tes mana yang diujikan. Jika tes yang diujikan mempunyai tingkat kesulitan tinggi, estimasi keampuan peserta tes akan rendah. Sebaliknya jika tes yang diujikan mempunyai tingkat kesulitan rendah, estimasi kemampuan peserta tes akan tinggi. Keterbatasan itulah yang kemudian mendorong para pakar psikologi dan pengukuran untuk mencari metode alternatif yang kemudian lahirlah teori respon butir.

Berbagai upaya terus dilakukan untuk mengembangkan alat tes yang fair dan objektif, mulai dari perbaikan teknik penulisan butir tes, adminsitrasi tes, sampai pengkajian dampak panjang tes dan ukuran sampel. Penelitian yang dilakukan oleh Hambleton dan Cook dengan menggunakan data simulasi, mengungkap pengaruh sampel size dan panjang tes terhadap kestabilan estimasi parameter butir dan kemampuan peserta tes pada *item respon theory* baik untuk 1, 2, ataupun 3 parameter.

Dari penelitian ini, Hambleton dan Cook membuat dua kesimpulan utama yaitu: (1) panjang tes memiliki hubungan terhadap kestabilan estimasi parameter butir dan kemampuan peserta tes, yang dibuktikan oleh nilai korelasi Rank Spearman rata-rata 0,80. Penambahan jumlah butir dapat meningkatkan kestabilan estimasi parameter butir dan kemampuan; (2) Ukuran sampel tulisan memiliki hubungan dengan kestabilan estimasi parametere butir dan kamampuan peserta tes. Semakin besar jumlah sampel tulisan semakin stabil estimasi parameter item butir, sebaliknya semakin kecil jumlah sampel tulisan semakin berkurang pula kestabilan estimasi parameter butir dan kemampuan.

Berdasarkan uraian di atas tulisan ini mencoba mencari bukti empiris dengan mereplikasi tulisan Hambleton dan Cook tersebut dengan menggunakan panjang tes dan ukuran sampel yang berbeda. Keunggulan tulisan ini dari tulisan Hambleton dan Cook adalah pada sisi jumlah sampel yang direplikasi. Pada penelitian Hambelton dan Cook dilakukan dengan jumlah sampel 500 peserta tes untuk melihat kekekaran model, sedangkan dalam tulisan ini mengkaji pengaruh ukuran sampel terhadap kekekaran model estimasi dengan menggunakan tiga kelompok tiga jenis ukuran sampel (271, 605, dan 1.999). Secara sederhana tujuan yang hendak dicapai penelitian ini adalah untuk menguji pengaruh panjang tes dan ukuran sampel terhadap kestabilan estimasi parameter item dan parameter kemampuan pada teori respon butir *(item response theory).* Tulisan ini dibatasi pada aplikasi estimasi parameter butir (tingkat daya beda, tingkat kesukaran, dan terkaan semu) teori respon butir pada model penilaian acuan norma.

Kajian Teori

Banyak definisi mengenai pengukuran. Salah satu diantaranya adalah definisi yang dikemukakan Stevens pada tahun 1946 yang mengatakan bahwa pengukuran adalah pemberian numeral atau angka kepada objek atau kejadian dengan menggunakan aturan-aturan tertentu (Crocker & Algina, 1986). Definisi itu kemudian

disempurnakan dengan mengatakan bahwa yang diberi atribut numeral bukanlah objek itu sendiri melainkan sifat-sifat yang melekat pada objek itu.

Pengukuran psikologis menjadi bagian baku kurikulum peserta tes psikologi dan pendidikan setelah pada tahun 1904 Thorndike mempublikasikan bukunya yang berjudul *An Introduction to the Theory of Mental and sosial Measurementi* (Crocker & Algina, 1986). Teori-teori yang ada di buku tersebut kemudian disempurnakan oleh pakar-pakar pengukuran setelah itu. Kumpulan *body of knowledge* tersebut, yang kemudian populer dengan nama teori tes klasik (*classical test theory)*, memberikan dasar teori untuk pengembangan tes kecerdasan, tes prestasi, tes kepribadian dan tes psikologis yang lain.

1. Model Tes

Secara umum, bentuk-bentuk tes dapat diklasifikasikan ke dalam: (1) tes pilihan ganda; (2) tes benar-salah; (3) tes isian/ jawaban singkat; (4) tes menjodohkan; dan (5) tes uraian (Umar, et.al, 1998). Berbagai bentuk tes tersebut mempunyai keunggulan dan kelemahan. Tes yang baik harus terdiri dari butir-butir soal yang baik. Sebagai contoh, untuk tes acuan norma, pendapat yang ditulis Budiyono (2005) dapat digunakan sebagai acuan. Budiyono menulis bahwa butir soal yang baik harus mempunyai tingkat kesulitan yang memadai, daya pembeda yang baik, dan berfungsinya pengecoh.

Ebel & Frisbie menulis, butir tes yang yang baik harus dikembangkan sesuai tujuan tes, mengetahui prestasi, pemetakan, memotivasi (1989: 20-21). Selanjutnya Ebel & Frisbie (1989: 24) menulis, tes yang dikembangkan oleh guru, penulis, atau pakar pengukuran, harus mengacu kepada bagaimana skor akan diinterpretasikan, apakah berbasis penilaian acuan norma (*norm referenced*), penilaian acuan kriteria (*criterion referenced*), ataukah berbasis penilaian acuan domain (*domain referenced*). Setiap model penilaian membutuhkan kriteria yang berbeda, misalnya pada model penilaian acuan norma, dibutuhkan parameter butir (tingkat daya

beda, tingkat kesukaran, dan terkaan semu) yang seragam untuk peserta tes, sedangkan untuk penilaian acuan kriteria dan domain dapat menggunakan parameter butir yang bervariasi. Sebagaimana dijelaskan di latar belakang masalah, tulisan ini lebih difokuskan aplikasi teori tes respon butir pada model penilaian acuan norma.

Pengukuran menurut teori tes klasik mengandung beberapa kelemahan, antara lain sebagai berikut. Pertama, tingkat kesulitan butir soal didefinisikan sebagai proporsi tes yang menjawab benar pada suatu sampel atau kelompok peserta tes tertentu. Ini berarti bahwa indeks tingkat kesulitan butir tergantung kepada peserta tes yang dikenai butir soal tersebut dan sebaliknya kemampuan para peserta tes tergantung kepada apakah butir-butir soal mudah atau sulit. Kedua, indeks daya pembeda suatu butir soal, koefisien validitas, dan koefisien reliabilitas skor tes juga tergantung kepada kelompok peserta tes yang dikenai tes tersebut.

2. Asumsi pada Teori Tes Klasik

Ada beberapa asumsi dasar pada teori tes klasik. Asumsi-asumsi itu, antara lain: (1) skor yang diperoleh oleh peserta tes terdiri dari skor sebenarnya (*Skor asli*) dan kesalahan pengukuran; (2) nilai harapan skor yang diperoleh sama dengan nilai harapan dari skor sebenarnya; (3) skor sebenarnya dan kesalahan pengukuran tidak berkorelasi; (4) kesalahan pengukuran pada dua tes yang mengukur kemampuan yang sama tidak saling berkorelasi; dan (5) pada dua tes yang mengukur kemampuan yang sama, kesalahan pengukuran pada tes pertama tidak berkorelasi dengan skor sebenarnya pada tes kedua (Allen & Yen: 1979; Sumadi Suryabrata: 2000).

Menurut Dali S. Naga (1992), pada pengukuran berdasar teori tes klasik, tes yang sama yang dijawab oleh kelompok peserta tes yang sama menghasilkan karakteristik yang berbeda. Dengan kata lain, karakteristik butir soal dipengaruhi oleh peserta tes yang menempuh tes yang berbeda, maka ciri kelompok peserta itu pada umumnya berubah. Ini berarti, ciri-ciri kelompok peserta tes berubah jika mereka menempuh tes yang berbeda.

3. Teori Respons Butir

Untuk mengatasi kelemahan-kelemahan yang ada pada teorites klasik,para ahli pengukuran berusaha mencari modedl alternatif. Model yang diinginkan harus mempunyai sifat-sifat: (1) karakteristik butir soal tidak tergantung kepada kelompok peserta tes yang dikenai butir soal tersebut; (2) skor yang menyatakan kemampuan peserta tes tergantung kepada tes; (3) model dinyatakan dalam tingkatan (*level*) butir soal, tidak dalam tingkatan tes; (4) model tidak memerlukan tes paralele untuk menghitung koefisien reliabilitas; dan (5) model menyediakan ukuran yang tepat untuk setiap skor kemampuan (Hambleton, et al, 1991). Model alternatif yang dapat mempunyai ciri-ciri itu adalah model pengukuran yang disebut teori respons butir (*item response theory*).

Ada tiga asumsi dasar yang mendasari teori respons butir, yaitu: (1) unidimensionalitas; (2) independensi lokal; dan (3) fungsi karakteristik butir menyatakan hubungan yang sebenarnya antara variabel yang tak terobservasi, yaitu kemampuan, dengan variabel terobsesi, yaitu respons butir. (Hambleton, et al, 1991; Sumadi Suryabrata, 2000). Asumsi unidimensionalitas dan independensi lokal dapat dijelaskan sebagai berikut.

Asumsi unidimensionalitas menyatakan bahwa hanya satu kemampuan yang diukur oleh sekumpulan butir-butir soal dalam suatu tes. Asumsi ini pada praktik sukar dipenuhi, sebab terdapat banyak faktor yang dapat mempengaruhi suatu tes kinerja. Faktor-faktor tersebut antara lain tingkatmotivasi, kecemasan, kemampuan untuk bekerja cepat, dan keterampilan kognitif lain di luar kemampuan yang diukur oleh sekumpulan yang diukur oleh sekumpulan butir soal. Hal yang dimaksud dengan unidimensionalitas ini adalah adanya faktor-faktor dominan yang mempengaruhi suatu tes kinerja. Faktor-faktor dominan itulah yang disebut kemampuan yang diukur oleh suatu tes.

Asumsi independensi lokal menyatakan bahwa jika kemampuan yang mempengaruhi suatu tes kinerja adalah konstan, maka respons peserta tes pada setiap pasangan butir soal adalah independen secara

statistik. Dengan kata lain, asumsi independensi lokal menyatakan bahwa tidak ada korelasi antara respons peserta tes pada butirsoal yang berbeda. Hal ini juga berarti bahwa kemampuan yang dinyatakan dalam model adalah satu-satunya faktor yang mempengaruhi respons peserta tes pada butir-butir soal.

Ada tiga model yang populer pada teori respons butir, yang cocok untuk tes dikotomous (termasuk tes pilihan ganda), yang disebut model logistik satu para meter (tingkat kesulitan = b), model logistik dua parameter (daya beda dan tingkat kesulitan = a, b), dan model logistik tiga parameter (daya beda, tingkat kesulitan, terkaan = a, b, c).

Parameter tingkat kesulitan, yaitu b, untuk sebuah butir soal adalah titik pada skala kemampuan, yang pada titik itu peluang menjawab benar butir tersebut sebesar 0,5 demikian pula peluang menjawab salah sebesar 0,5 (Hambleton, et al, 1991). Jika kemampuan (θ) ditransformasikan demikian hingga mempunyai rerata (*rerata)* 0 dan simpangan baku 1, maka nilai b biasanya berkisar antara -2 sampai dengan 2 (Hambleton, et al, 1991). Butir soal yang tingkat kesulitannya mendekati -2 merupakan butir soal yang sangat mudah dan butir soal yang tingkat kesulitannya mendekati 2 merupakan butir soal yang sangat sukar.

Parameter daya pembeda, yaitu a, proporsional terhadap koefisien arah garis singgung (*slope*) pada titik θ = b (Hambleton, Swaminathan, Rogers, 1991). Butir soal yang mempunyai daya pembeda yang besar mempunyai kurva yang sangat menanjak, sedangkan butir soal yang mempunyai daya pembeda yang kecil mempunyai nilai mulai dari $-\infty$ sampai dengan $+\infty$. Namun demikian, untuk butir soal yang baik, nilai parameter a harus terletak antara 0 dan 2 (Hambleton, et al, 1991).

4. Penelitian Monte Carlo

Penelitian dengan data simulasi yang dilakukan oleh Hambleton dan Cook menguji pengaruh panjang tes dan ukuran sampel terhadap kestabilan estimasi parameter butir dan kemampuan peserta tes.

Simulasi model teori respon butir dilakukan dengan menggunakan model 3 parameter. Pertama kali ditentukan jumlah peserta tes (N), bentuk distribusi kemampuan, nilai parameter kemampuan. Kedua menentukan jumlah butir tes dan nilai parameter a, b, dan c. Ketiga mensubstitusikan paratemer item dengan peserta tes dalam persamaan logistik 3 parameter untuk mendapatkan p_u ($0 \leq p_u \leq 1$), yang menggambarkan probabilitas peserta tes ke i menjawab benar item ke j.

Tulisan Hambleton dan Cook ini menggunakan ukuran sampel 500 peserta tes dengan menggunakan model tiga parameter. Data parameter butir (a, b, c) berdistribusi uniform dan parameter kemampuan peserta tes berdistribusi normal, keduanya dibangkitkan dengan program aplikasi DATA GENERATE

Harwell, dkk (1996) menulis materi sejenis dengan tulisan Hambleton dan Cook. Harwell dkk. menjabarkan berbagai hal berhubungan dengan tulisan Monte Carlo (RMC) untuk mengestimasi parameter butir dan kemampuan pada teori respon butir. Pada *review* teorinya, Harwell mengutip asumsi penggunaan RMC dari Majalah *Policy of Psychometrika (Psychometric Society)* yang diterbitkan tahun 1979, bahwa: (1) data mustahil dikumpulkan; (2) sulit menentukan teknik sampling yang akan digunakan dalam analisis; dan (3) perbandingan alogaritma tersedia bagi analisis fungsi tertentu.

Di samping itu, ada alasan lain yaitu apabila tulisan atau eksperimen itu berdampak membahayakan khalayak, misalnya dampak penambahan panas tabung gas terhadap daya ledak. Keunggulan tulisan RMC yang paling utama adalah RMC dapat dilakukan walaupun problem praktis tidak dijumpai atau ada problem praktis namun sangat kompleks untuk dipecahkan secara prakmatik. Di samping itu, dengan RMC peneliti dapat memanipulasi parameter atau asumsi dan meneliti pengaruh beberapa variable sekaligus. Namun demikian RMC memiliki keterbatasan yaitu manfaat tulisan sangat dipengaruhi oleh seberapa realistis model yang dikembangkan oleh peneliti, kualitas data yang

digunakan sulit untuk dinilai, dan hasil sangat bervariasi pada setiap replikasi dan tingkat presisi program komputer yang digunakan.

RMC dalam teori respon butir pada umumnya bertujuan untuk menguji prosedur estimasi parameter, menguji asumsi-asumsi statistik dalam teori respon butir, dan membandingkan metode yang digunakan dalam teori respon butir.

Metode Analisis

Wright dan Stone (dalam Whitmore & Schumacker, 1999: 917) merekomendasikan ukuran sampel minimum 200 pada metode pendeteksian *DIF* berdasarkan model Rasch. Ini berarti, ukuran sampel minimum 200 dapat dipakai jika digunakan metode pendeteksian *DIF* berdasarkan model yang hanya memperhatikan tingkat kesulitan (b). Karena tulisan ini membandingkan juga metode estimasi yang berdasarkan model logistik tiga parameter, maka menurut Drasgow (Robie, Mueller, & Champion, 2001: 181) diperlukan ukuran sampel yang lebih besar daripada 200, walaupun Camilli dan Shepard (1994: 157) mengatakan bahwa *"even for sampel size of only a few hundred examinees, IRT models may be preferable"*.

Untuk metode regresi logistik, Rogers dan Swaminathan (1993: 107) menggunakan ukuran sampel 250 dan 500 per kelompok. Penelitian Hambleton dan Rogers (1989: 318) menggunakan 1000 siswa *Anglo American* dan 1000 siswa *Native American.* Kim dan Cohen (1995: 294) menggunakan 606 siswa yang menggunakan kalkulator dan 624 siswa yang tidak menggunakan kalkulator. Swaminathan dan Rogers (1990) menggunakan ukuran 250 dan 500 untuk masing-masing kelompok acuan dan kelompok fokus.

Berdasarkan pendapat pakar-pakar tersebut di atas, paling tidak dapat diklasifikasi menjadi tiga kisaran golongan jumlah sampel , yaitu 250, 600, dan 1000. Untuk mengetahui pengaruh jumlah sampel tersebut dalam penelitian ini di ukuran sampel yang mendekati ukuran sampel yang digunakan pada penelitian terdahulu yaitu, 271, 605, dan 1.999. Penentuan jumlah sampel sebesar itu

dimaksudkan untuk mengetahui pola perubahan kekekaran estimasi fungsi butir pada setiap penambahan ukuran sampel sebagaimana dilakukan oleh Budiyono (2005).

Menurut Mislevy dan Bock (1990: 1-10), tes dengan panjang 11-20 butir merupakan tes pendek dan tes dengan panjang lebih dari 20 butir disebut tes panjang. Berdasarkan kepada dua hal itu, pada tulisan ini, dilihat dari panjangnya, tes dibedakan atas dua macam, yaitu: (1) tes dengan panjang 15 butir (mewakili tes pendek); dan (2) tes dengan panjang 55 butir (mewakili tes panjang).

Data tulisan ini merupakan data hipotetis yang diperoleh dengan menggunakan program aplikasi DGEN yang dikembangkan oleh R. K. Hambleton dan R. J. Rovinelli dari universitas Massachusetts dan dikembangkan lagi oleh HJ Rogers dari Universitas Colombia pada tahun 1992. Parameter butir (tingkat daya beda = a, tingkat kesulitan = b, pseudo terkaan = c, dan tingkat ability = θ), ukuran sampel, dan panjang tes ditentukan awal. Secara sederhana range nilai parameter dan distribusi, ukuran sampel, dan panjang tes disajikan dalam table berikut ini.

Tabel 1. Rancangan Panjang Test dan Ukuran Sampel

Panjang tes	**Ukuran sampel**	**a**	**b**	**c**	**θ**
15	271	0.40 - 1.90	-2.00 - 2.00	0.00 - 0.15	0.00 - 1.00
	605				
	1999				
55	271	0.40 - 1.90	-2.00 - 2.00	0.00 - 0.15	0.00 - 1.00
	605				
	1999				
Distribusi		**Uniform**	**Uniform**	**Uniform**	**Normal**

Parameter skor asli yang dibangkitkan dengan program DGEN meliputi nilai tingkat daya beda (a), tingkat kesulitan (b), tingkat *pseudoguessing* (c), dan kemampuan (θ). Program BILOG digunakan untuk mengestimasi parameter skor asli untuk setiap panjang tes (15 item dan 55 item) dan setiap ukuran sampel (271, 605, 1999)

diulang sampai 5 kali, sehingga total klasifikasi data yang akan dikorelasikan ada 120 kelompok.

Menurut Harwell dkk. (1996) dan Harwell (1997) tahapan yang dilalui dalam RMC ada tiga, yaitu tahap perumusan masalah, tahap mengembangkan desain tulisan, tahap mengidentifikasi dan memvalidasi program komputer yang digunakan untuk menghasilkan data (*generating seed number to get data*), dan menganalisis hasil.

Teknik analisis yang digunakan meliputi analisis tabulasi, grafik, teknik korelasi *product moment*, rerata galat kuadrat (MSE), galat baku (SE), analisis fungsi informasi, serta analisis multivariat. Berdasarkan analisis MSE dan SE, sebuah estimasi parameter dikatakan lebih stabil apabila nilai MSE dan SE lebih kecil nilainya. Sebaliknya berdasarkan analisis korelasi, apabila nilai korelasi estimasi parameter dengan skor asli lebih tinggi dapat disimpulkan bahwa estimasi parameter tersebut lebih stabil. Analisis table dan grafik digunakan untuk menyederhanakan pola hubungan antara nilai estimasi parameter dengan nilai skor asli baik dari sudut pandang MSE, SE, maupun korelasinya. Analisis multivariate digunakan untuk menguji pengaruh panjang tes dan ukuran sampel secara bersama terhadap kestabilan estimasi parameter butir dan kemampuan.

Hasil Analisis

Berdasarkan hasil pengolahan data dengan program aplikasi DEGEN dan BILOG diperoleh data skor asli dan skor estimasi untuk setiap panjang tes dan setiap ukuran sampel sejumlah 120 kelompok, dengan rincian sebagai berikut.

Tabel 2. Jumlah Replikasi Data Tulisan

Panjang tes	Ukuran sampel	a	B	c	θ	Program
15	271	5	5	5	5	DGEN
	605	5	5	5	5	
	1999	5	5	5	5	
55	271	5	5	5	5	
	605	5	5	5	5	
	1999	5	5	5	5	
Jumlah		30	30	30	30	120
Panjang tes	**Ukuran sampel**	**a**	**B**	**c**	**θ**	
15	271	5	5	5	5	BILOG
	605	5	5	5	5	
	1999	5	5	5	5	
55	271	5	5	5	5	
	605	5	5	5	5	
	1999	5	5	5	5	
Jumlah		30	30	30	30	120
Total		60	60	60	60	240

Berdasarkan hasil estimasi yang dilakukan dengan program BILOG diperoleh nilai estimasi parameter kemampuan, daya beda, tingkat kesulitan, dan terkaan. Selanjutnya nilai estimasi dan skor asli parameter digunakan untuk menentukan nilai korelasi, rerata galat kuadrat, galat baku, fungsi informasi dan nilai statistic multivariate. Nilai-nilai statistik itu kemudian disajikan dalam bentuk grafik untuk melihat gerakan nilai untuk setiap panjang tes dan ukuran sampel.

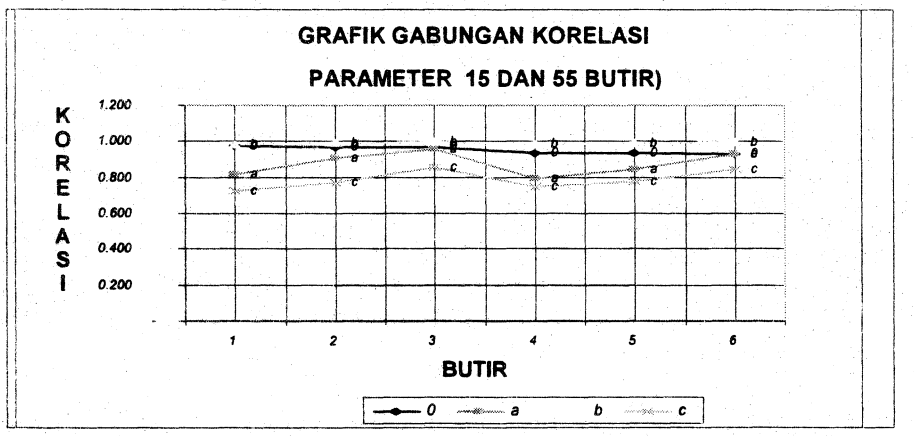
1. Kestabilan Estimasi Berdasar Skor Korelasi

Pada grafik 1 berikut ini ditunjukkan pergerakan nilai korelasi pada setiap panjang tes dan ukuran sampel. Nilai korelasi estimasi

parameter kemampuan dan tingkat kesulitan antara skor asli dengan skor estimasi tidak menunjukkan perubahan berarti walaupun panjang tes dan ukuran sampel ditambah. Namun demikian korelasi keduanya menunjukkan nilai tertinggi apabila dibandingkan dengan parameter tingkat daya beda, dan terkaan.

Variasi nilai korelasi parameter tingkat daya beda sebelum panjang tes ditambah dengan setelah ditambah tidak begitu jauh berbeda. Penambahan jumlah sampel dapat meningkatkan kestabilan estimasi parameter yang dibuktikan oleh kenaikan grafik parameter tingkat daya beda dan terkaan. Pada grafik a dan c berikut ini tampak nilai awal korelasi pada panjang tes 55 saat ukuran sampel 271 dan 605 lebih tinggi daripada skor korelasi pada panjang tes 15 saat ukuran sampel 271 dan 605. Dengan demikian dapat disimpulkan panjang tes dan ukuran sampel berpengaruh terhadap kestabilan estimasi parameter, semakin panjang tes dan besar ukuran sampel semakin stabil estimasi parameter. Skor korelasi pada kedua panjang tes (15 dan 55 butir) serta ketiga ukuran sampel 271, 605 dan 1.999, berkisar antara 0,70 sampai dengan 0,99.

Grafik 1. Gabungan Korelasi Item (15 dan 55 butir)



Keterangan:

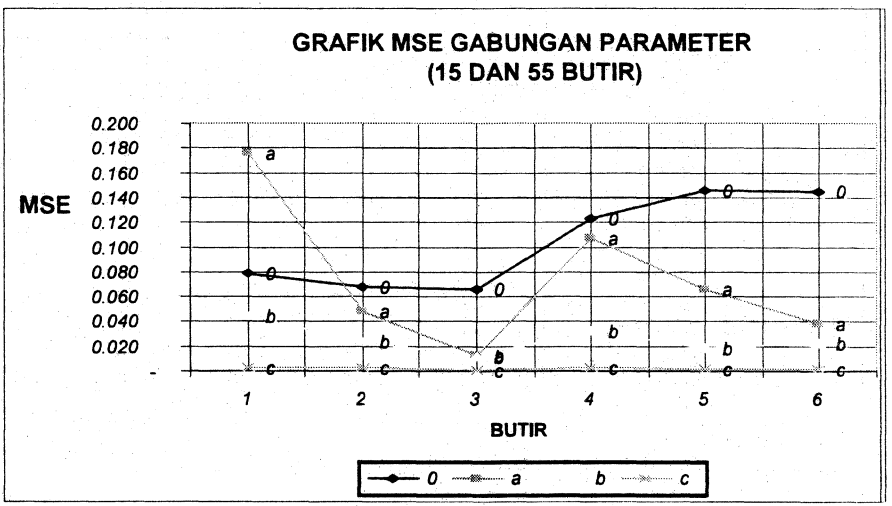
1 – 3 : Jumlah butir 55 dengan ukuran sampel berturutan 271, 605, 1999

4 – 6 : Jumlah butir 15 dengan ukuran sampel berturutan 271, 605, 1999

2. Kestabilan Estimasi Berdasar Skor MSE

Grafik MSE gabungan berikut ini menggambarkan pergerakan nilai MSE parameter dari panjang tes 15 ke 55 butir dan ukuran sampel 271 sampai ke 1.999. Grafik MSE tingkat daya beda panjang tes 15 butir turun sangat tajam setelah jumlah sampel ditambah, bahkan pada saat jumlah sampel 1.999, skor MSE panjang tes 15 butir lebih rendah daripada panjang tes 55 pada jumlah sampel yang sama. Namun demikian secara umum, penambahan panjang tes dan jumlah sampel akan meningkatkan kestabilan estimasi butir.

Grafik 2. Gabungan MSE Item (15 dan 55 butir)



Grafik MSE parameter tingkat kesulitan bervariasi seperti parameter tingkat daya beda, namun variasi MSE antarukuran sampel dan panjang tes tidak sebesar yang terjadi pada parameter tingkat daya beda. Selanjutnya nilai MSE parameter terkaan tidak mengalami perubahan signifikan walaupun panjang tes dan ukuran sampel diubah. Sebaliknya nilai MSE parameter kemampuan justru

cenderung bergerak naik, baik karena panjang tes ditambah ataupun jumlah sampel ditambah.

3. Kestabilan Estimasi Berdasar Analisis Multivariat

Untuk melihat pengaruh secara bersama panjang tes dan ukuran sampel terhadap kestabilan estimasi parameter butir (a, b, c) dan untuk memperjelas grafik pengaruh yang telah disajikan terdahulu, maka pada bagian berikut ini disajikan uji statistiknya.

Tabel 3. Hasil Analisis Variate Panjang Tes dan Ukuran Sampel terhadap Kestabilan Estimasi Parameter Butir

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	,916	841,473[a]	3,000	232,000	,000
	Wilks' Lambda	,084	841,473[a]	3,000	232,000	,000
	Hotelling's Trace	10,881	841,473[a]	3,000	232,000	,000
	Roy's Largest Root	10,881	841,473[a]	3,000	232,000	,000
PTES	Pillai's Trace	,000	.[a]	,000	,000	.
	Wilks' Lambda	1,000	.[a]	,000	233,000	.
	Hotelling's Trace	,000	.[a]	,000	2,000	.
	Roy's Largest Root	,000	,000[a]	3,000	231,000	1,000
UKSAM	Pillai's Trace	,000	.[a]	,000	,000	.
	Wilks' Lambda	1,000	.[a]	,000	233,000	.
	Hotelling's Trace	,000	.[a]	,000	2,000	.
	Roy's Largest Root	,000	,000[a]	3,000	231,000	1,000
PTES * UKSAM	Pillai's Trace	,000	.[a]	,000	,000	.
	Wilks' Lambda	1,000	.[a]	,000	233,000	.
	Hotelling's Trace	,000	.[a]	,000	2,000	.
	Roy's Largest Root	,000	,000[a]	3,000	231,000	1,000

a. Exact statistic

b. Design: Intercept+PTES+UKSAM+PTES * UKSAM

Berdasarkan uji statistik multivariate dengan SPSS 11.5 diperoleh nilai Hotelling's Trace sebesar 10,881 dengan tingkat signifikansi 0,000. Nilai ini membuktikan panjang tes dan besar sampel berpengaruh terhadap estimasi parameter butir (a, b, c).

Tabel 2 berikut ini menunjukkan nilai F hitung sebesar 0,000 dengan tingkat signifikansi 1,000. Angka ini menunjukkan bahwa

panjang tes dan ukuran sampel tidak memiliki pengaruh terhadap estimasi parameter kemampuan.

Tabel 4. Hasil Analisis Multivariate Panjang Tes dan Ukuran Sampel terhadap Kestabilan Estimasi Kemampuan

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6,185E-11[a]	2	3,092E-11	,000	1,000
Intercept	,000	1	,000	,000	1,000
PATES	,000	0	.	.	.
UKSAMP	,000	0	.	.	.
PATES * UKSAMP	,000	0	.	.	.
Error	5329,330	5747	,927		
Total	5329,330	5750			
Corrected Total	5329,330	5749			

a. R Squared = ,000 (Adjusted R Squared = ,000)

4. Kestabilan Estimasi Berdasar Galat Baku dan Fungsi Informasi

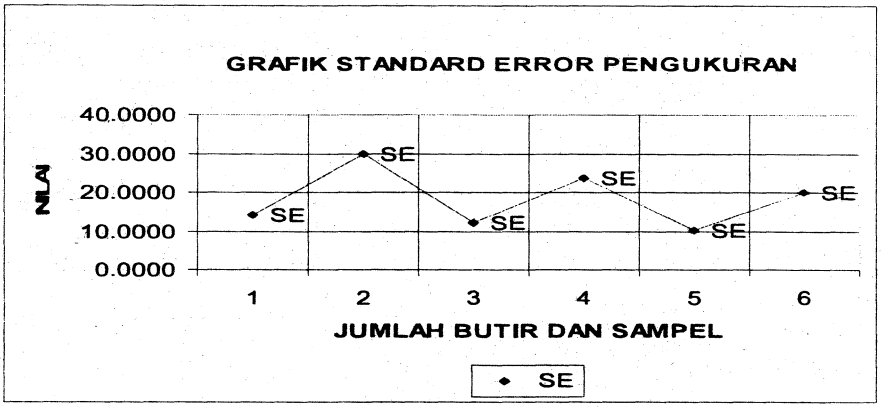
Tabel 5. Galat baku dan Fungsi Informasi

Keterangan	Nilai	Keterangan	Nilai
Rerata SE 15-271	14.1620	FI 15-271	0.8866
Rerata SE 15-605	29.8060	FI 15-605	0.3753
Rerata SE 15-1999	12.2880	FI 15-1999	1.2615
Rerata SE 55-271	23.6100	FI 55-271	0.6807
Rerata SE 55-605	10.2370	FI 55-605	1.3215
Rerata SE 55-1999	20.1180	FI 55-1999	1.0784

Keterangan : Rerata SE = Rerata galat baku (SE)
FI = Fungsi informasi

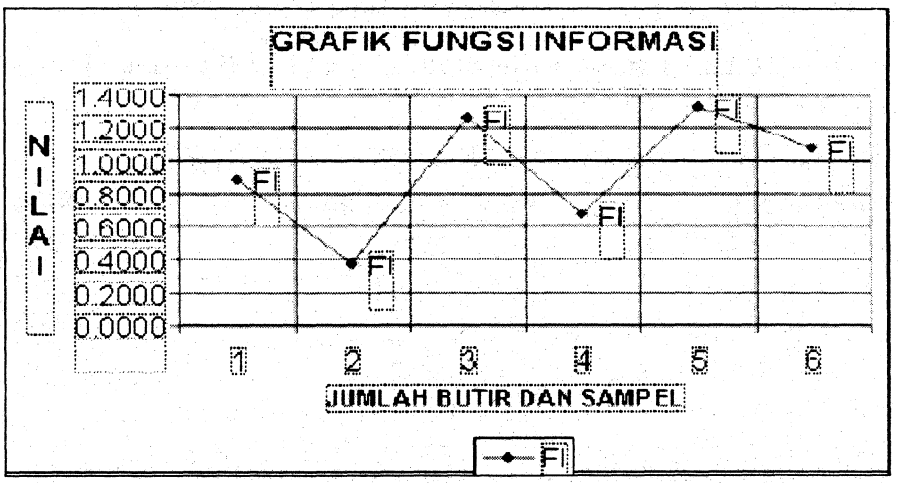
Angka galat baku dan fungsi informasi di atas secara sederhana disajikan pada grafik berikut ini.

Grafik 3. Galat Baku (15 dan 55 butir)



Keterangan :
1 – 3 : jumlah butir 15 dengan jumlah sampel 271, 605, 1.999
4 – 6 : jumlah butir 55 dengan jumlah sampel 271, 605, 1.999

Grafik 4. Fungsi Informasi (15 dan 55 butir)



Keterangan :
1 – 3 : jumlah butir 15 dengan jumlah sampel 271, 605, 1.999
4 – 6 : jumlah butir 55 dengan jumlah sampel 271, 605, 1.999

Rerata galat baku menunjukkan besaran galat estimasi parameter dari skor aslinya. Model estimasi yang baik, seharusnya memiliki galat baku yang kecil. Sedangkan fungsi informasi menunjukkan seberapa baik nilai informasi mencerminkan skor asli. Semakin tinggi fungsi informasi semakin baik suatu model estimasi. Dari table 8 di atas, terlihat pada saat jumlah butir 55 dan jumlah sampel 605, nilai standar error terkecil (10.2370), sedangkan pada saat panjang tes 15 butir dan besar sampel 605, nilai standar error terbesar (29.8060). Hasil tersebut konsisten apabila dilihat dari nilai fungsi informasinya. Pada saat panjang tes dan jumlah sampel memiliki nilai informasi terbesar (1.3215), sedangkan saat panjang tes dan besar sampel nilai fungsi informasi terkecil (0.3753).

Kesimpulan

Berdasarkan analisis data dengan berbagai teknik (grafik, matematik, statistik), dapat disimpulkan tiga hal berikut.

1. Ukuran sampel berpengaruh terhadap kestabilan estimasi parameter butir daya beda dan terkaan, tetapi tidak berpengaruh terhadap parameter tingkat kesulitan.
2. Penambahan ukuran sampel dan penambahan panjang tes akan mengurangi nilai MSE dan SE pada estimasi parameter butir parameter butir daya beda dan tingkat kesulitan, tetapi tidak berpengaruh terhadap parameter terkaan.
3. Panjang tes tidak berpengaruh terhadap nilai fungsi informasi sedangkan ukuran sampel berpengaruh terhadap nilai fungsi informasi.

Keterbatasan

1. Tidak memberikan aksioma tentang variable panjang dan besar ukuran sampel dalam kaitannya dengan formula 1, 2, dan 3 parameter pada model teori respon butir.

2. Artikel ini menggunakan data simulasi, sehingga data itu dapat saja tidak sesuai dengan keadaan sesungguhnya di lapangan. Apabila hal itu terjadi, berarti hasil tulisan ini akan menyesatkan.
3. Data simulasi sangat dipengaruhi oleh setting ketelitian computer, misalnya berapa dua angka decimal yang diset dalam computer akan mempengaruhi ketelitian hasil.
4. Data simulasi kadang-kadang dikembangkan bukan dari permasalahan praktis, sehingga kadang hasilnya tidak dapat diterapkan atau tidak dijumpai di lapangan (*external validity* rendah).

Saran

1. Tulisan dengan data simulasi seperti ini dapat direplikasi kembali oleh para peneliti di Indonesia sehingga diperoleh kesimpulan yang lebih meyakinkan tentang perlu tidaknya menambah atau mengurangi panjang tes dan ukuran sampel pada saat mengembangkan tes. Hal itu sangat penting dilakukan karena kesalahan penentuan panjang tes dan ukuran sampel memiliki dampak yang sangat besar terhadap penilaian hasil belajar pada tingkat sekolah hingga tingkat nasional.
2. Tulisan ini dapat dikembangkan dengan menggunakan variable outlier, anchor tes, DIF, Equiting, atau bahkan pemodelan parameter apabila variabelnya ditambah menjadi 4 parameter, 5, 6 dan seterusnya.
3. Tulisan demikian juga dapat dikombinasi dengan menggunakan data real. Data riil tersebut digunakan untuk menentukan pola distribusi dan pola respon sebagai dasar untuk membangkitkan data simulasi, sehingga hasil penelitian bukan merupakan produk murni simulasi tetapi sudah disesuaikan dengan lapangan.

Daftar Pustaka

Allen, M.J. & Yen, M.W. 1979. *Introduction to Measurement Theory*. Monterey : Brook / Cole Publishing Company.

Azwar, S. 1999. *Dasar-dasar Psikometri*. Yogyakarta : Pustaka Pelajar.

Budiyono. 2005. "Perbandingan Metode Mantel – Haenszel, Sibtest, Regresi Logistik, dan Perbedaan Peluang Dalam Mendeteksi Keberbedaan Fungsi Butir". *Disertasi*: Pascasarjana Universitas Negeri Yogyakarta.

Camilli, S. & Shepard, L.A. 1994. *Methods for Identifying Biased Test Items*. Thousand Oaks, CA : Sage Publications.

Crocker, L. & Algina J. 1986. *Introduction to Classical and Modern Test Theory*. New York : CBS College Publishing.

Ebel, R. L. & David A. F. 1989. *Essentials of Educational Measurement*. Fourth Edition. New Jersey: Prentice Hall Inc.

Hambleton, R. K. & Linda L. C. 1983. "Robustness of Item Response Models and Effects of Tes Length and Sampel Size on the Precision of Ability Estimates". *New Horizons in Testing*. New York : Academic Press.

Hambleton, R.K. et al. 1991. "Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel – Haenszel Methods". *Applied Measurement in Education*.

Harwell, M. R. 1997. *Analyzing the Results of Monte Carlo Tulisanes in Item Response Theory*. London : Sage Publication Inc.

Harwell, M., dkk. 1996. "Monte Carlo Tulisanes in Item Response Theory". *Applied Psychological Measurement*. Vol. 20. No. 2 June 1996, pp. 101 – 115.

Jahja, U., et al. 1998. *Bahan Penataran Pengujian Pendidikan*. Jakarta: Pusat Tulisan dan Pengembangan Pendidikan, Balitbangdikbud, Departemen Pendidikan dan Kebudayaan.

Linn , R. L. (Ed). 1989. *Educational Measurement*. New York : Macmillan Publishing Company.

Mardapi, D. 2004. "Penyusunan Tes Hasil Belajar". *Makalah*, Tidak Dipublikasikan. Yogyakarta: Program Pasca Sarjana Universitas Negeri Yogyakarta.

MCDonald. 1999. Test Theory: *A Unified Treatment*. Mahwah. NJ: Lawrence Erlbaum Associates.

Mislevy, R. J dan Bock, R. D. 1990. *Bilog 3. Item Analysis and Test Scoring with Binary Logistic Models*. Second Edition. USA: Scientific Software Inc.

Naga, D. S. 1992. *Pengantar Teori Skor Pada Pengukuran Pendidikan*. Jakarta : Gunadarma.

Olmedilla, J. M. M. 1992. "Tradition and Change in National Examination Sistems: A Comparison of Mediterranean and Anglo Saxon Countries". *Examinations: Comparative and International Tulisanes*. Oxford: Pergamon Press.

Robie, C. M, & Campion, J.E. 2001. "Effects of a Motivational Inducement on the Psychometric Properties of Cognitive Ability Test". *Journal of Business and Psychology*. 2. pp. 177 – 189.

Suryabrata, S. 2000. *Pengembangan Alat Ukur Psikologis*. Yogyakarta: Andi Offset.

Swediati, N. 1997. "Equating Tests Under The Generalized Partial Credit Model". *Dissertation* Presented by School of Education in University of Massachusetts Amherst.

Whitmore, M.L. & Schumaker, R.E. 1999. A Comparison of Logistic Regression and Analysis of Variance Differential Item Functioning Detection Methods. *Educational and Psychological Measurement*, 59. pp. 910 – 927.