PENGARUH *SAMPLE SIZE* (*N*) DAN *TEST LENGTH* (*n*) TERHADAP *ITEM PARAMETER ESTIMATE* DAN *EXAMINEE PARAMETER ESTIMATE*, SUATU STUDI SIMULASI

R. BUDIARTI[1]

Abstrak

Studi yang mempelajari masalah pengukuran secara umum di bidang pendidikan dan mempelajari metode untuk menyelesaikannya, telah berkembang menjadi disiplin ilmu khusus yang dikenal dengan *test theory*. *Test theory* menyediakan kerangka kerja umum untuk melihat proses pembentukan instrumen tes (*item test*). Analisis item dapat dilakukan dengan pendekatan tes teori klasik (*Classical Test Theory* atau *CTT*) dan teori tes modern yang dikenal dengan *Item Respons Theory* (*IRT*). Ada beberapa model respon item (*item response model*), yang berbeda banyaknya parameter dalam model. Semua model *IRT* mengandung satu atau lebih parameter item dan satu atau lebih parameter *examinee*. Pada tulisan ini difokuskan pada model respon item dengan satu parameter *examinee* dengan dua parameter item. Parameter-parameter ini tidak diketahui, untuk itu perlu diduga. Agar hasil dugaan relatif stabil dan akurat, maka diperlukan *sample size* yang cukup. Tujuan dari paper ini adalah (1) menginvestigasi pengaruh *sample size* (*N*) terhadap kestabilan *item parameter estimate*, (2) menginvestigasi pengaruh *test length* (*n*) terhadap kestabilan *examinee parameter estimate*. Kestabilan dugaan parameter item (*a* dan *b*) dipengaruhi oleh *sample size*, dan kestabilan parameter *examinee* (θ) dipengaruhi oleh ukuran *test length*. Semakin besar *sample size*, maka pendugaan parameter item makin stabil, sedangkan semakin besar ukuran *test length* maka makin stabil dugaan parameter item.

PENDAHULUAN

Studi yang mempelajari masalah pengukuran secara umum di bidang pendidikan dan bidang psikologi dan mempelajari metode untuk menyelesaikannya, telah berkembang menjadi disiplin ilmu khusus yang dikenal dengan *test theory*. Metode yang dipelajari dalam *test theory* meliputi (1) metode untuk menduga sejauh mana masalah ini mempengaruhi pengukuran yang diambil dalam situasi tertentu, (2) merumuskan metode untuk mengatasi atau meminimumkan masalah ini (Crocker dan Algina, 1986). *Test theory* menyediakan kerangka kerja umum untuk melihat proses pembentukan instrumen tes (*item test*). Analisis item dapat dilakukan dengan pendekatan tes teori klasik

[1]Departemen Matematika, Fakultas Ilmu Pengetahuan Alam, Jalan Meranti Kampus IPB Dramaga Bogor, 16680.

(*Classical Test Theory* atau *CTT*) dan teori tes modern yang dikenal dengan *Item Respons Theory* (*IRT*).

Item Respons Theory (*IRT*) berlandaskan pada dua postulat dasar yaitu (1) kinerja dari peserta ujian (*examinee performance*) pada *test item* dapat diprediksi (atau dapat diterangkan) melalui himpunan faktor-faktor disebut 'kemampuan' (*trait, latent trait*, atau *ability*), dan (2) hubungan antara kinerja item peserta ujian (*examinees' item*) dengan himpunan dari kemampuan yang berlandaskan pada kinerja item dapat digambarkan oleh fungsi monoton naik yang disebut *item characteristic function* atau *item characteristic curve* (*ICC*). Peserta ujian dengan nilai kemampuan yang besar berarti bahwa peserta tersebut mempunyai peluang besar untuk dapat menjawab item dengan benar, sebaliknya peserta ujian dengan nilai kemampuan yang kecil berarti bahwa peserta tersebut mempunyai peluang kecil untuk dapat menjawab item dengan benar. *Item characteristic function* merupakan fungsi monoton naik, artinya jika tingkat kemampuan peserta ujian meningkat maka peluang menjawab item dengan benar juga meningkat.

Ada beberapa model respon item (*item response model*), yang berbeda bentuk matematika dari *item characteristic function* dan berbeda banyaknya parameter dalam model. Semua model *IRT* mengandung satu atau lebih parameter item dan satu atau lebih parameter *examinee*. Pada tulisan ini difokuskan pada model respon item dengan satu parameter *examinee* dengan dua parameter item. Parameter-parameter ini tidak diketahui, untuk itu perlu diduga. Agar hasil dugaan relatif stabil dan akurat, maka diperlukan *sample size* yang cukup.

Pertanyaan tentang kecukupan *sample size* (banyaknya *examinee*) sering muncul. Pertanyaan ini muncul pada diskusi sejumlah topik, termasuk diskusi tentang apakah tersedia literatur yang berkaitan dengan rekomendasi mengenai *sample size*. Menurut Crocker dan Algina (1986), bahwa tidak ada aturan mutlak mengenai *sample size* minimum yang digunakan dalam studi analisis item, Crocker dan Algina juga menyatakan bahwa *sample size* yang dibutuhkan tergantung pada pemilihan model tertentu. Menurut Xing dan Hambleton (2002) bahwa secara umum, makin panjang *test length* (n) menghasilkan reliabilitas tinggi. Kualitas item yang bagus akan meningkatkan reliabilitas, sedangkan kualitas item yang buruk akan mengurangi reliabilitas.

Pertanyaan penting selanjutnya adalah seberapa dekat hubungan antara *true scores* (parameter *examinee* atau parameter item) dan *observed scores*? Indeks hubungan ini adalah koefisien korelasi antar dua variabel tersebut. Koefisien korelasi yang menunjukkan derajat hubungan antara *true scores* dan *observed scores* dikenal dengan indeks reliabilitas (*reliability index*), dan koefisien korelasi ini disebut juga dengan koefisien stabilitas (*coefficient of stability*), (Crocker dan Algina, 1986). Semakin tinggi indeks reliabilitas maka *observed scores* semakin mirip dengan *true scores*, dengan kata lain, nilai dugaan semakin stabil mendekati nilai parameter yang sebenarnya.

Jadi permasalahannya adalah bagaimanakah pengaruh *sample size* (N), *test length* (n) dan model respon item terhadap dugaan parameter *examinee* (*examinee*

parameter estimate) dan dugaan parameter item (*item parameter estimate*). Berdasarkan permasalahan ini, maka tujuan dari paper ini adalah

(1) menginvestigasi pengaruh *sample size* (N) terhadap kestabilan *item parameter estimate,*

(2) menginvestigasi pengaruh *test length* (n) terhadap kestabilan *examinee parameter estimate.*

METODE

Distribusi *Latent Trait* (Sebaran Parameter Kemampuan/*Ability*)

Seperti dituliskan pada judul paper ini adalah suatu studi simulasi, ditentukan simulasi sampel *latent trait* (θ) berasal populasi normal baku ($\theta \sim N(0,1)$), seperti yang dilakukan Linn, Levine, Hastings, dan Wadrop (1981). Berikut ini dituliskan beberapa definisi yang dibutuhkan untuk pembahasan lebih lanjut.

Item Response Model

Item characteristic function adalah ekspresi matematika yang menghubungkan antara peluang menjawab benar item, untuk mengukur kemampuan peserta tes (*examinee*), dan karakteristik item. Sementara itu ada tak terhingga banyaknya model IRT, hanya beberapa yang digunakan. Asumsi yang mendasari semua model IRT adalah hanya ada satu parameter kemampuan (θ), sehingga seringkali disebut model IRT unidimensional. Perbedaan utama antar model IRT unidimensional adalah banyaknya parameter yang digunakan untuk menggambarkan item-item. Tiga model IRT unidimensional yang paling dikenal adalah model logistik satu-parameter, dua-parameter dan tiga-parameter. Model-model ini sesuai untuk data respon item dikotomus (Hambleton *et. al.*, 1991).

Model Logistik satu-parameter (model 1*PL*)

Model logistik satu-parameter (model 1PL) adalah satu dari model IRT yang paling banyak digunakan. Model 1PL sering juga disebut model Rasch (Rasch, 1960). *Item characteristic curve* untuk model logistik satu-parameter diberikan oleh persamaan berikut :

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad ; i = 1, 2, \ldots, n$$

dengan

$P_i(\theta)$: peluang *examinee* dengan kemampuan θ dapat menjawab item ke-i dengan benar

b_i : parameter tingkat kesulitan item ke-i

n : banyaknya item dalam suatu tes

e : bilangan transendental yaitu 2.718.

Ketika nilai kemampuan (θ) suatu grup *examinee* ditransformasi sehingga rata-ratanya sama dengan 0 dan standar deviasinya sama dengan 1, nilai-nilai b_i cenderung bervariasi diantara -2 dan 2. Nilai b_i yang dekat dengan -2 berarti bahwa item sangat mudah, sebaliknya jika nilai b_i dekat dengan 2 berarti bahwa item sangat sulit bagi grup *examinee* tersebut.

Asumsi yang mendasari model 1PL (selain unidimensional) adalah tingkat kesulitan item merupakan satu-satunya karakter item yang mempengaruhi kinerja *examinee* (*examinee performance*). Hal ini berarti bahwa semua item mempunyai tingkat pembeda yang sama dan ICC mempunyai *lower asymptote* bernilai 0 (artinya bahwa peluang *examinee* memiliki tingkat kemampuan sangat rendah mendekati 0).

Model Logistik dua-parameter (model 2PL)

Lord (1952) adalah orang pertama yang memkonstruksi model respon item dua-parameter yang berdasarkan pada sebaran normal kumulatif (normal ogive). Birnbaum (1968) menyubstitusi model logistik dua-parameter (model 2PL) dari fungsi ogive normal dua-parameter sebagai bentuk fungsi karakteristik item. Fungsi logistik memiliki keuntungan dalam praktek dibandingkan dengan fungsi ogive normal, karena fungsi ogive mengadung bentuk integral. *Item characteristic curve* untuk model logistik dua-parameter ditemukan oleh Birmbaum, yang diberikan oleh persamaan berikut :

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad ; i = 1, 2, \ldots, n$$

dengan

$P_i(\theta)$: peluang *examinee* dengan kemampuan θ dapat menjawab item ke-i dengan benar

b_i : parameter tingkat kesulitan item ke-i

n : banyaknya item dalam suatu tes

e : bilangan transendental yaitu 2.718.

D : faktor skala sehingga membuat fungsi logistik menjadi sedekat mungkin dengan fungsi ogive normal (D = 1.7)

a_i : parameter pembeda item ke-i

Secara teori, parameter pembeda (a_i) didefinisikan pada interval $(-\infty, \infty)$. Item-item pembeda bernilai negatif dibuang dari tes kemampuan, berarti ada yang salah dari item-item tersebut karena peluang menjawab benar turun saat kemampuan *examinee* naik. Juga, tidak umum nilai a_i mencapai lebih besar dari 2.

Umumnya, nilai parameter pembeda a_i berkisar pada interval $(0,2)$, (Hambleton *et al*, 1991).

Seperti model 1PL, asumsi yang mendasari model 2PL (selain unidimensional) adalah tingkat kesulitan item dan tingkat pembeda merupakan karakter item yang mempengaruhi kinerja *examinee* (*examinee performance*). Hal ini berarti bahwa semua item mempunyai ICC dengan *lower asymptote* bernilai 0 (artinya bahwa peluang *examinee* memiliki tingkat kemampuan sangat rendah mendekati 0).

Model Logistik tiga-parameter (model 3PL)

Ekspresi matematik untuk model logistik tiga-parameter diberikan oleh persamaan berikut :

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \quad ;i=1,2,\ldots,n$$

dengan

$P_i(\theta)$: peluang *examinee* dengan kemampuan θ dapat menjawab item ke-*i* dengan benar

b_i : parameter tingkat kesulitan item ke-*i*

n : banyaknya item dalam suatu tes

e : bilangan transendental yaitu 2.718.

D : faktor skala sehingga membuat fungsi logistik menjadi sedekat mungkin dengan fungsi ogive normal ($D = 1.7$)

a_i : parameter pembeda item ke-*i*

c_i : parameter menebak (*guessing*) item ke-*i*

Parameter menebak (c_i) disebut juga dengan parameter *pseudo-chance-level.* Parameter c_i menetapkan *nonzero lower asymptote* pada kurva karakteristik item dan merepresentasikan peluang *examinee* dengan kemampuan rendah menjawab item dengan benar (Hambleton *et al*, 1991).

Sample Size

Kecukupan *sample size* (*N*) menjadi topik perdebatan, berikut ini beberapa literatur yang merekomendasikan pemilihan *sample size.*

Menurut Crocker dan Algina (1986), secara umum parameter item dapat diduga dengan relatif stabil untuk sampel sebanyak 200 *examinee.* Berdasarkan *rule-of-thumb* (Nunnally, 1967) besarnya *sample size* adalah 5 sampai 10 kali banyaknya item. Misalkan suatu tes terdiri dari 20 item, maka minimal *size sample* sebanyak 100 *examinee.* Crocker dan Algina (1986) merekomendasikan bahwa *sample size* yang dibutuhkan untuk pendugaan parameter berdasarkan teori respon item bervariasi antara 200 sampai dengan 1000 subjek. Jika digunakan model 3PL, Lord (1968) merekomendasikan banyaknya item $n \geq 50$ dan *sample*

size sebesar $N \geq 1000$. Banyak peneliti telah mereferensikan seperti Lord (1968) dan Hulin *et al* (1982) bahwa *sample size* yang direkomendasikan minimal 1000 *examinee* untuk kalibrasi model 3PL. Berdasarkan studi sebelumnya (Hulin *et al*, 1982) dinyatakan bahwa banyaknya item $n = 50$ dan *sample size* $N = 1000$ sudah dianggap cukup besar untuk mendapatkan pendugaan parameter item yang akurat, ketika asumsi unidimensional dipenuhi.

Ukuran Kestabilan Dugaan

Ketika peneliti memberikan suatu tes, mereka hanya mengetahui *observed score*. Pertanyaan penting adalah seberapa dekat hubungan antara *true score* (*examinees' score*) dengan *observed score*? Satu indeks hubungan ini adalah korelasi antara kedua variabel tersebut. Koefisien korelasi yang mengekspresikan tingkat hubungan antara *true* dan *observed score* pada suatu tes dikenal sebagai *reliability index*. Mengingat kembali *examinee's observed score* diekspresikan sebagai berikut :

$$X = T + E$$

Dan dalam *deviation score*, ditulis

$$x = t + e$$

Ketika menggunakan *deviation score*, *reliability index* dapat diekspresikan sebagai berikut :

$$\rho_{XT} = \frac{\sum xt}{N\sigma_X\sigma_T}$$

dengan

N : *sample size*

σ_X : simpangan baku *observed score* (nilai dugaan)

σ_T : simpangan baku *true score* (nilai parameter)

Menurut Crocker dan Algina (1986), koefisien korelasi ini dikenal juga sebagai *coefficient of stability*. Oleh karena itu, koefisien korelasi di atas dapat digunakan sebagai ukuran kestabilan dari dugaan suatu parameter.

Selain menggunakan koefisien korelasi, Lord dan Novick (1968) menyatakan bahwa ukuran kestabilan dapat juga menggunakan *root mean squared differences* (*RMSD*). *RMSD* untuk parameter *a, b, c* dan parameter θ ditulis:

$$RMSD(a) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{a}_i - a_i)^2}$$

$$RMSD(b) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{b}_i - b_i)^2}$$

$$RMSD(c) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{c}_i - c_i)^2}$$

$$RMSD(\theta) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}$$

Berdasarkan rumus *RMSD* di atas, dapat diinterpretasikan bahwa jika selisih antara *true score* (nilai parameter) dan *observed score* (nilai dugaan) kecil, artinya dugaannya akurat, maka nilai *RMSD* nya kecil juga.

Untuk menjawab tujuan (1) dan (2) pada paper ini, digunakan model respon item dua-parameter (model 2PL). Pada model ini, peluang *examinee* memberikan respon item ke-*i* dengan *latent trait* (θ) unidimensional tertentu, seperti persamaan model 2PL yang ditulis sebelumnya di atas. Berdasarkan alasan–alasan yang dikemukakan sebelumnya, simulasi parameter a_i ditentukan berdistribusi uniform (0.4, 2) dan parameter b_i ditentukan berdistribusi uniform (-2, 2). Model-model IRT unidimensional, termasuk model 2PL, sesuai untuk data respon item dikotomus (Hambleton *et. al.*, 1991), sehingga dalam simulasi ditentukan respon item dikotomus.

***Test Length* (*n*) dan *Sample Size* (*N*)**

Berdasarkan alasan secara teori maupun berdasarkan penelitian sebelumnya yang telah dikemukakan, maka ditentukan n dan N untuk masing-masing tujuan sebagai berikut :

(1) Untuk menjawab tujuan (1) dan mengacu pada *rule-of-thumb*, ditentukan $n = 40$ dan $N = 200$, 400, dan 1000.

(2) Untuk menjawab tujuan (2) dan mengacu pada *rule-of-thumb*, ditentukan $N = 1000$ dan $n = 20$, 50, dan 100.

Ditentukan replikasi/ulangan sebanyak 10 kali, dan untuk mengukur kestabilan dugaan parameter digunakan indeks reliabilitas dan RMSD (*root mean squared differences*).

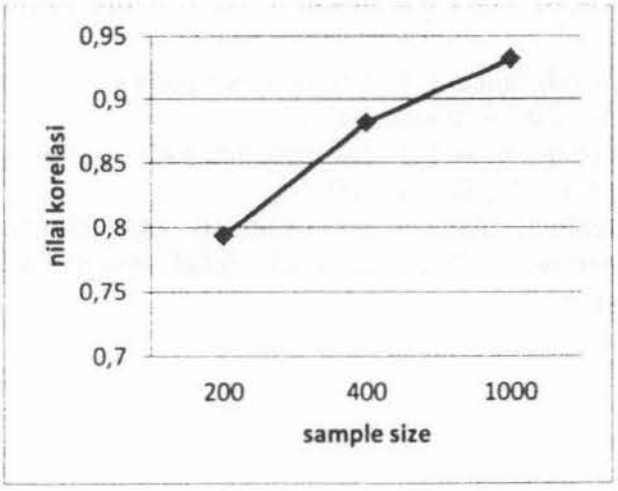
PEMBAHASAN

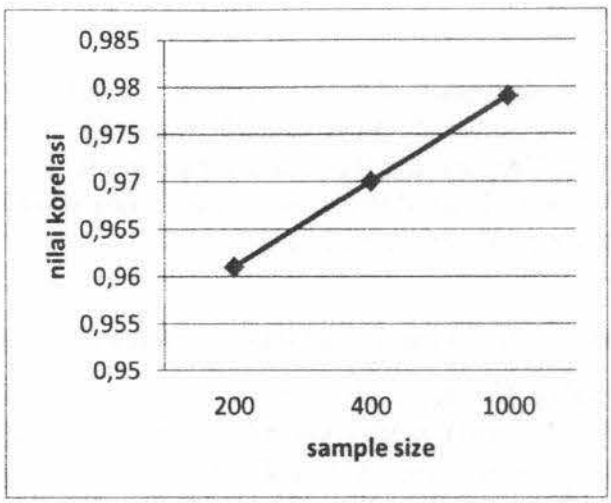
Pengaruh *sample size* (*N*) terhadap kestabilan *item parameter estimate*

Untuk mengetahui pengaruh *sample size* (*N*) terhadap kestabilan dugaan parameter item *a* dan *b* pada model logistik dua-parameter (2PL), ditetapkan banyaknya item (*test length*) $n = 40$ dan *sample size* dibuat bervariasi yaitu $N =$ 200, 400, 1000, masing-masing variasi ini direplikasi sebanyak 10 kali. Sudah disebutkan sebelumnya, untuk melihat kestabilan dugaan parameter item *a* dan *b* digunakan koefisien korelasi dan RMSD (*root mean squared differences*) atau RMSE (*root mean squared error*). Hasil korelasi dan RMSD dari perameter item *a* dan *b* dapat dilihat pada Tabel 1 berikut dan diperjelas dengan menampilkannya dalam bentuk grafik.

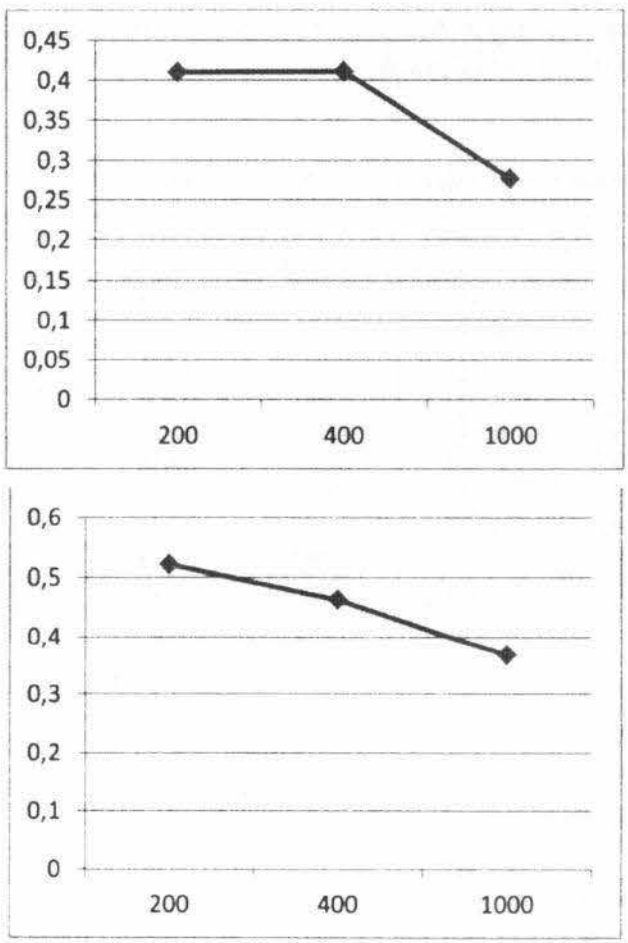
TABEL 1
Hasil korelasi dan RMSD parameter item *a* dan *b*

Replikasi	Korelasi parameter *a*			Korelasi parameter *b*			RMSD parameter *a*			RMSD parameter *b*		
	200	**400**	**1000**	**200**	**400**	**1000**	**200**	**400**	**1000**	**200**	**400**	**1000**
1	0.793	0.881	0.933	0.974	0.975	0.976	0.521	0.436	0.307	0.477	0.430	0.395
2	0.833	0.911	0.940	0.954	0.975	0.979	0.469	0.377	0.294	0.578	0.431	0.378
3	0.770	0.784	0.918	0.955	0.967	0.979	0.447	0.421	0.301	0.535	0.412	0.357
4	0.818	0.858	0.937	0.973	0.955	0.972	0.443	0.426	0.330	0.444	0.474	0.400
5	0.776	0.881	0.955	0.973	0.972	0.976	0.410	0.411	0.277	0.459	0.391	0.396
6	0.710	0.841	0.886	0.947	0.959	0.979	0.406	0.406	0.320	0.521	0.463	0.368
7	0.855	0.925	0.918	0.944	0.974	0.981	0.442	0.342	0.322	0.613	0.446	0.373
8	0.814	0.835	0.945	0.966	0.959	0.981	0.415	0.457	0.323	0.544	0.471	0.367
9	0.787	0.877	0.922	0.956	0.977	0.978	0.508	0.497	0.325	0.510	0.404	0.384
10	0.844	0.761	0.940	0.975	0.959	0.965	0.507	0.461	0.316	0.404	0.446	0.442





Gambar 1. Korelasi antara *true score a* dengan dugaan *a* (gambar atas) dan korelasi antara *true score b* dengan dugaan *b* (gambar bawah)



Gambar 2. RMSD parameter item *a* (gambar atas) dan RMSD parameter item *b* (gambar bawah)

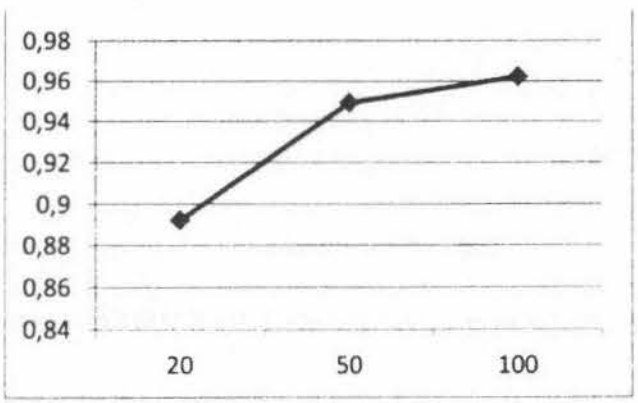
Tabel 1 yang diperjelas dengan Gambar 1 dan Gambar 2 di atas memperlihatkan bahwa semakin besar *sample size* (N) maka nilai korelasi semakin besar juga, sedangkan nilai RMSD semakin kecil. Jadi *sample size* (N) berpengaruh terhadap kestabilan dugaan parameter, yaitu semakin besar *sample size* (N) maka dugaan parameter semakin stabil.

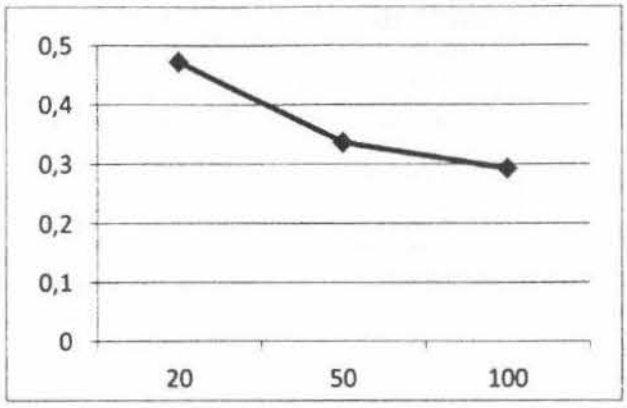
Pengaruh *test length* (n) terhadap kestabilan *examinee parameter estimate*

Untuk mengetahui pengaruh *test length* (n) terhadap kestabilan dugaan parameter *examinee* (θ) pada model logistik dua-parameter (2PL), ditetapkan banyaknya *examinee* (*sample size*) N = 1000 dan *test length* dibuat bervariasi yaitu n = 20, 50, 100, masing-masing variasi ini direplikasi sebanyak 10 kali. Sudah disebutkan sebelumnya, untuk melihat kestabilan dugaan parameter *examinee* (θ) digunakan koefisien korelasi dan RMSD (*root mean squared differences*) atau RMSE (*root mean squared error*). Hasil korelasi dan RMSD dari perameter *examinee* dapat dilihat pada Tabel 2 berikut dan diperjelas dengan menampilkannya dalam bentuk grafik.

TABEL 2
Hasil korelasi dan RMSD parameter *examinee* (θ)

Replikasi	Korelasi			RMSD		
	n = 20	n = 50	n = 100	n = 20	n = 50	n = 100
1	0.893	0.951	0.963	0.463	0.320	0.278
2	0.892	0.949	0.96	0.466	0.326	0.291
3	0.892	0.946	0.963	0.465	0.334	0.278
4	0.895	0.945	0.962	0.461	0.336	0.285
5	0.889	0.946	0.960	0.472	0.336	0.293
6	0.897	0.950	0.962	0.456	0.322	0.282
7	0.896	0.949	0.962	0.457	0.324	0.285
8	0.890	0.952	0.962	0.470	0.317	0.282
9	0.888	0.950	0.962	0.473	0.323	0.285
10	0.890	0.952	0.960	0.470	0.317	0.290





Gambar 3. Korelasi parameter *examinee* (θ) (gambar atas) dan RMSD parameter *examinee* (θ) (gambar bawah)

Tabel 2 di atas yang diperjelas dengan Gambar 3 menunjukkan bahwa semakin besar *test length* (n) maka nilai korelasi semakin besar juga, sedangkan nilai RMSD semakin kecil. Nilai korelasi antara "true" parameter examinee (θ) dengan nilai dugaannya hamper mendekati 1 pada $n = 50$ dan $n = 100$. Hal ini berarti bahwa untuk keperluan membentuk instrumen tes dapat digunakan $n = 50$ atau $n = 100$. Jika mempertimbangkan biaya, maka dapat digunakan $n = 50$. Jadi *test length* (n) berpengaruh terhadap kestabilan dugaan parameter *examinee* (θ), yaitu semakin besar *test length* (n) maka dugaan parameter *examinee* (θ) semakin stabil.

SIMPULAN

Kestabilan dugaan parameter item (a dan b) dipengaruhi oleh *sample size*, dan kestabilan parameter *examinee* (θ) dipengaruhi oleh ukuran *test length*. Semakin besar *sample size*, maka pendugaan parameter item makin stabil, sedangkan semakin besar ukuran *test length* maka makin stabil dugaan parameter item.

DAFTAR PUSTAKA

[1] Crocker, L dan Algina, J. 1986. *Introduction to classical and modern test theory*. Rinehart and Winston, Inc. Amerika Serikat.

[2] Drasgow, F dan Parsons, CK. 1983 . Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*. Vol. 7 : No. 2, pp 189-199.

[3] Hambleton, RK, Swaminathan, H dan Rogers, HJ. 1991. *Fundamentals of item response theory*. Sage Publication, California.

[4] Hullin CL, Lissak RI, Drasgow F. 1982. Recovery of two- and three-parameter logistic item characteristic curve. A monte carlo study. *Applied Psychological Measurement*. Vol. 7 : No. 6, pp. 249-260.

[5] Linn RL, Levine MV, Hastings CN, dan Wardrop JL. 1981. Item bias in a test of reading comprehension. *Applied Psychological Measurement*, 5, 159-173.

[6] Lord FM. 1968. An analiysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989-1020.

[7] Lord FM dan Novick MR. 1968. *Statistical theories of mental test scores*. Reading MA : Addison-Wesley.