ESTIMASI *TRUE SCORE* PADA *SECOND ORDER* UNIDIMENSIONAL DATA: SEBUAH STUDI SIMULASI MONTE CARLO TENTANG DAMPAK PANJANG TES, TINGKAT KESUKARAN DAN DAYA PEMBEDA ITEM

Puti Febrayosi
Fakultas Psikologi UIN Jakarta

Abstrak:

Tujuan penelitian ini adalah untuk mengetahui apakah terdapat perbedaan atau bias pada data unidimensional second order namun sering diperlakukan sebagai unidimensional hanya pada tingkat pertama atau first order, serta mengetahui sejauh mana pengaruh panjang tes, heterogenitas tingkat kesukaran dan daya pembeda apabila terdapat perbedaan atau bias terhadap kemampuan responden. Penelitian ini merupakan penelitian simulasi Monte Carlo dengan 27 model percobaan dan setiap model direplikasi sebanyak 50 kali. Model second order unidimensional yang dibangkitkan memiliki panjang tes 20, 40 dan 60 item, dengan heterogenitas tingkat kesukaran dan daya pembeda 0,025, 0,10 dan 0, 20. Software komputer yang digunakan adalah Mplus, dengan bantuan estimator Bayesian*. Untuk mengetahui apakah terdapat perbedaan atau bias antara first order dan second order maka yang dilihat nilai mean yang dihasilkan lebih besar dari nol. Hasil penelitian ini menunjukan: (1) semua model unidimensional pada* second order *namun dianalisis sebagai unidimensional pada* first order *hasil yang diperoleh mengenai theta (θ atau kemampuan responden) tidak memberikan gambaran yang sebenarnya, karena terdapat bias atau perbedaan dari nilai mean yang dihasilkan lebih besar dari nol; (2) bias atau perbedaan dari theta (θ atau kemampuan responden) paling besar dihasilkan oleh panjang tes 20 item dengan daya pembeda 0.20 dan tingkat kesukaran 0,10 sedangkan bias atau perbedaan dari theta (θ atau kemampuan responden) paling kecil dihasilkan oleh panjang tes 60 item dengan daya pembeda dan tingkat kesukaran 0,20; (3) disamping itu, berdasarkan hasil perhitungan diperoleh R square sebesar 0.130, hal ini berarti 13% bias responden dapat dijelaskan oleh bervariasinya*

panjang tes, heterogenitas tingkat kesukaran dan daya pembeda dengan indeks signifikansi sebesar 0.007 (p < 0.05). Dengan demikian apabila data unidimensional pada second order namun menganalisisnya hanya pada first order unidimensional *menghasilkan bias serta tidak memberikan gambaran seutuhnya mengenai kemampuan responden. Kalaupun tetap memberlakukan* first order *pada data unidimensional second order, bias paling kecil diperoleh dengan panjang tes yang lebih besar. Dalam penelitian ini tes dengan panjang 60 item bias yang dihasilkan lebih rendah dibandingkan tes dengan panjang 20 item.*

Kata Kunci: *second order unidimensional*, panjang tes, tingkat kesukaran, daya pembeda, dan monte carlo

Pendahuluan

Secara sederhana, tes didefinisikan sebagai alat ukur atau prosedur (Ronald, 2010), sedangkan pengetesan (Kaplan, 1993) diartikan sebagai pengukuran atau teknik yang digunakan untuk mengukur perilaku atau membantu memahami dan memprediksi perilaku. Pengetesan baik di bidang pendidikan ataupun psikologi memiliki tujuan tertentu seperti menempatkan seseorang pada tempat yang tepat sesuai dengan bidangnya, menjadi bahan pertimbangan untuk kebijakan yang akan diambil, atau bahan evaluasi proses belajar mengajar. Untuk itu tes yang digunakan harus memiliki kualitas item yang baik dan berkualitas tinggi. Analisis item bertujuan untuk mengidentifikasi mana item-tem yang baik, kurang ataupun tidak baik sama sekali, sehingga ketika digunakan hasil tes tersebut benar-benar sudah mengukur apa yang hendak diukur atau yang ingin diketahui. Analisis terhadap kualitas item dilakukan baik secara kualitatif maupun kuantitatif. Analisis item secara kualitatif ialah mengkaji secara teoritik item tes yang telah disusun, dengan memperhatikan tiga aspek, yaitu aspek materi, aspek konstruksi, dan aspek bahasa. Sedangkan analisis item secara kuantitatif dapat menggunakan pendekatan teori tes klasik (*classical test theory)* maupun teori respon item *(item response theory).*

Teori tes klasik merupakan sebuah teori yang sudah digunakan dalam kurun waktu yang lama, sehingga sebagian besar orang yang terkait dengan dunia pendidikan dan psikologi telah mengetahui dan memahami konsep serta penerapan-nya. Salah satu keunggulan dari tes ini terletak pada konsepnya yang sederhana untuk menghitung koefisien validitas dan reliabilitas tes, parameter soal dan kemudahan menentukan kemampuan peserta. Skor amatan

dalam teori tes klasik (*observed score* = X) terdiri dari skor sebenarnya (*true score* = T) dan skor kesalahan (*error score* = E). Nilai *true score* merupakan nilai rata-rata yang diperoleh dari pengulangan tes menggunakan soal tes yang sama, dan menentukan kemampuan peserta tes dengan cara menjumlahkan skor amatan yang diperoleh peserta. Hal ini dapat dilakukan apabila item-item di dalamnya memiliki tingkat kesukaran dan daya pembeda nilai yang sama serta uni-dimensional. Jika kondisi di atas dapat terpenuhi, maka skor pada item-item tersebut dapat langsung dihitung dengan menjumlahkan semua skor pada item tersebut artinya skor total atau skor mentah tanpa pembobotan (Umar, 2012). Namun, kenyataannya *unidimensional test* sulit terpenuhi karena tingkat kesukaran dan daya pembeda yang bervariasi. Apabila ini digunakan maka dapat menimbulkan kerugian bagi pemakai hasil tes tersebut, terlebih lagi jika digunakan untuk mengambil sebuah keputusan. Maka keputusan tersebut menjadi kurang valid, hasilnya bias, makin besar penyimpangannya dan pemanfaatan *raw score* pada tes klasik menjadi kurang bermanfaat.

Untuk mengatasi kelemahan teori tes klasik, maka berkembanglah *item response theory* (IRT). Teori ini berkembang sangat pesat, tidak hanya pada bidang pendidikan dan psikologi, namun digunakan juga pada rekruitmen dan seleksi (misal, penerimaan pegawai atau mahasiswa baru), *qualification testing* (meng-kualifikasikan seseorang sesuai pada level-level tertentu), evaluasi program dan *assessment*, bidang klinis serta metode pengukuran dan penelitian. IRT digunakan secara luas dalam pengembangan tes, analisis dan seleksi item, penyetaraan tes, analisis bias item sampai dengan tes adaptif secara komputer atau *computerized adaptive test*, CAT (du Toit, 2003).

Pendekatan teori tes klasik dan IRT memiliki sudut pandang yang berbeda, tes klasik lebih berorientasi kepada test secara keseluruhan sedangkan IRT memfokuskan pada item IRT (pola jawaban responden). Menurut Hambleton (1991) keung-gulan yang dimiliki IRT antara lain: (a) karakteristik item tidak tergantung pada responden; (b) nilai kemampuan responden tidak tergantung pada tes yang dikerjakan; (c) model lebih menekankan tingkatan (*level*) butir soal daripada tingkatan tes; (d) tidak memerlukan tes paralel untuk menghitung koefisien realibilitas; dan (e) model menyediakan ukuran yang tepat untuk setiap skor kemampuan.

IRT memiliki dua postulat (Hambleton, 1991) yakni (a) performa dari responden dapat diprediksi atau dijelaskan oleh sekumpulan faktor yang disebut dengan *traits*, *laten traits* atau kemampuan (b) hubungan antara performa responden dalam item dengan performa responden dalam *traits* dapat dijelaskan melalui fungsi yang disebut dengan *item cha-racteristic function* atau *item characteristic curve* (ICC). Fungsi ini

menggambarkan bahwa semakin tinggi kemampuan seseorang maka semakin besar kemungkinan atau peluang seseorang untuk menjawab benar item tersebut.

Model yang digunakan dalam *item characteristic function* atau *item characteristic curve* (ICC) merupakan persamaan matematika yang menggambarkan hubungan antara kemungkinan jawaban yang benar dan kemampuan responden. Pada mulanya bentuk penyelesaian ICC menggunakan model kurva normal, namun karena sulitnya penghitungan maka digunakanlah bentuk kurva logistik.

Model logistik yang digunakan untuk data dikotomi dikenal dengan sebutan model satu, dua dan tiga parameter logistik. Perbedaan nama ini dikarenakan jumlah parameter yang digunakan didalamnya yaitu tingkat kesukaran, daya beda dan *pseudo guessing*. Sedangkan untuk data politomi terdapat beberapa model yaitu *partial credit model* (PCM), *graded response model* (GRM), dan *generalized partial credit model* (GPCM). Namun, penelitian kali ini hanya memfokuskan pada pola respon dikotomi dengan menggunakan model dua parameter logistik.

Model logistik dalam IRT memiliki persamaan bentuk umum (Crocker & Aligna, 1986) yakni:

$$P_g\theta = 1 + x^{x}$$

di mana e adalah dasar dari sistem natural logaritma, x merupakan *arbitrary symbol* (bukan menunjukan skor yang teramati). Perbedaan ketiga model tersebut tergantung dari banyaknya parameter yang digunakan untuk menggambarkan karakteristik item dalam model.

Model satu parameter logistik merupakan model yang sering digunakan dalam IRT. Model ini disebut dengan model satu parameter karena hanya terdapat satu parameter item didalamnya yakni tingkat kesukaran item yang dinotasikan dengan huruf "b". Jadi, kemungkinan jawaban benar responden hanya ditentukan oleh tingkat kesukaran item, sedangkan daya beda dianggap sama untuk semua item dalam sebuah tes. Dalam Hambleton, Swaminathan & Rogers (1991) persamaan model 1 parameter logistik sebagai berikut:

$$\theta = 1 + e^{\theta - b_i} \qquad i = 1, 2, \ldots n$$

Keterangan:

= probabilitas bagi responden dengan kemampuan (θ) untuk dapat menjawab item ke-*i* dengan benar

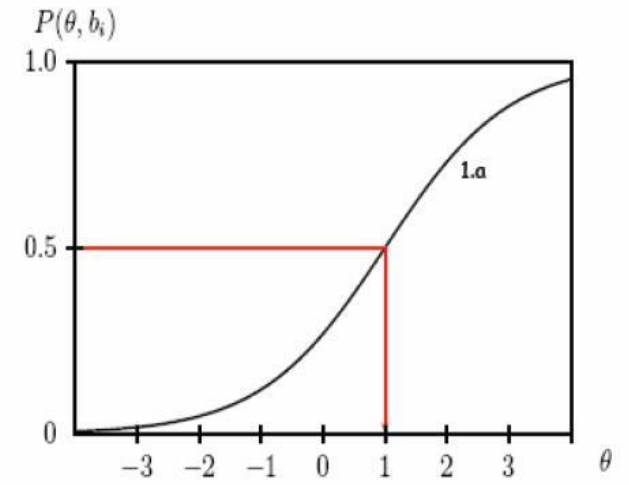
= parameter tingkat kesukaran item *i*

= jumlah item dalam tes

e = nilai transedental (eksponen) sebesar 2.718

= berbentuk kurva "huruf S" yang memiliki nilai antara 0 dan 1.

Di bawah ini merupakan gambar kurva karakteristik item model 1PL



Gambar 2.1. Kurva karakteristik item model 1 parameter logistik

Parameter b_i untuk item adalah titik pada skala kemampuan dimana probabilitas atau kemungkinan respon yang benar sebesar 0.5. Parameter ini menunjukkan posisi ICC dalam kaitannya dengan skala kemampuan. Semakin besar nilai parameter b_i, semakin besar kemampuan yang diperlukan responden untuk memiliki kesempatan 50% menjawab item dengan benar. Item dikatakan sulit apabila terletak di sebelah kanan atau lebih tinggi pada skala kemampuan; sedangkan item yang mudah berada di sebelah kiri atau di bawah pada skala kemampuan. Ketika nilai kemampuan dari kelompok diubah, sedemikian sehingga nilai rata-rata menjadi 0 dan standar deviasi menjadi 1 (satu), maka nilai b_i menjadi lebih bervariasi yang (biasanya) berada pada interval -2 sampai dengan +2. Nilai b_i dekat -2.0 maka termasuk item yang sangat mudah, dan nilai-nilai bi dekat +2.0 termasuk item yang sangat sulit.

Pada model 2 parameter logistik, kemungkinan responden untuk menjawab benar ditentukan oleh dua parameter yakni tingkat kesukaran dan daya pembeda. Setiap item memiliki daya beda yang berbeda-beda. Dalam Hambleton, Swaminathan & Rogers (1991) apabila terdapat item dengan daya pembeda besar maka kurva yang ditampilkan akan menanjak tajam, dibandingkan item dengan daya pembeda kecil, yang kurvanya akan lebih landai. Secara teoritis, nilai parameter daya pembeda terletak diantara -∞ dan +∞, namun efektif pada nilai 0 hingga 2. Model 2PL dikembangkan oleh Lord (dalam Hambleton, 1991) berdasarkan distribusi normal kumulatif (normal ogive). Kemudian, Birnbaum (dalam

Hambleton, 1991) mengusulkan model dua parameter yang menggunakan item kurva karakteristik item dengan fungsi distribusi logistik:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

$i = 1, 2, \ldots n$

Keterangan:

$P_i(\theta)$ = probabilitas dari kemampuan responden (θ) yang dapat menjawab item ke-*i* dengan benar

a_i = parameter daya pembeda

b_i = parameter tingkat kesukaran item

n = jumlah item dalam tes

e = nilai transedental (eksponen) sebesar 2.718

D = faktor penskalaan sebesar 1.7

atau persamaan model 2 parameter yang dapat ditulis dengan cara yang lain, apabila pembilang dan penyebut dari persamaan di atas digantikan dengan $e^{-Da_i(\theta - b_i)}$, sehingga , menjadi:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

Atau ditulis lebih sederhana menjadi

$$P_i(\theta) = \frac{1}{1 + e^{-z_i}}$$

Birnbaum menggantikan fungsi distribusi dua parameter yang awalnya fungsi normal ogive menjadi logistik kumulatif dalam bentuk item kurva karakteristik. Kurva logistik memiliki keuntungan karena lebih mudah dihitung dari pada kurva normal ogive. Model logistik lebih "*mathematically tractable*" dari pada model normal ogive karena normal ogive melibatkan integrasi fungsi eksplisit dari parameter item dan kemampuan. Penafsiran P_i (θ), b_i, a_i dan θ pada dasarnya sama seperti pada penafsiran dalam model normal ogive. Nilai konstanta D merupakan faktor penyesuaian skala. Sehingga perbedaan antara nomal ogive dan logistik pada 2PL kurang dari 0.01 untuk semua nilai θ. Jadi apabila kita menggunakan normal ogive dan logistik tidak memberikan perbedaan yang berarti dan signifikan. Daya beda model 2 parameter dalam kurva karakteristik item disebut dengan *slope parameter*¸sedangkan tingkat kesukaran disebut dengan *location parameter*.

Model tiga parameter logistik dapat diperoleh dari model dua parameter dengan menambahkan parameter ketiga, dinotasikan c_i. Bentuk matematis dari kurva logistik tiga parameter ditulis

$$P_i(\theta) = c_i + (1 - c_i)\, P_2(\theta)$$

$i = 1, 2, \ldots n\backslash$

Keterangan:

$P_i(\theta)$ = probabilitas dari responden dengan kemampuan (θ) untuk dapat menjawab item ke-*i* dengan benar

a_i = parameter daya pembeda

= parameter tingkat kesukaran item
= parameter tebakan atau *pseudo guessing*
= jumlah item dalam tes

e = nilai transedental (eksponen) sebesar 2.718

D = faktor penskalaan sebesar 1.7

Dalam hal kurvakarakteristik item, parameter ini menyediakan asimptot lebih tinggi dari 0 (nol) dan mewakili probabilitas peserta ujian dengan kemampuan sangat rendah untuk menjawab item dengan benar. Parameter dimasukkan ke dalam model untuk memperhitungkan kemungkinan responden menebak dalam tes yang itemnya bersifat pilihan ganda. Biasanya, diasumsikan sebagai nilai yang lebih kecil dari nilai yang akan terjadi jika peserta ujian menebak secara acak pada item test. Lord (dalam Hambleton, Swaminathan & Rogers, 1991) mencatat, bahwa fenomena ini mungkin dapat dikaitkan dengan kecerdikan pembuat item dalam mengembangkan pilihan (*distractor* atau pengecoh) yang menarik untuk dipilih tetapi tidak merupakan jawaban benar. Untuk alasan seperti ini, tidak boleh disebut "parameter menebak atau *guessing*".

Penggunaan model dan parameter item yang berbeda, akan menghasilkan estimasi kemampuan orang yang berbeda. Dalam IRT, tidak hanya parameter item yang akan mempengaruhi hasil estimasi kemampuan peserta tes (Lord & Novick dalam Ching-Fung, 2002), tetapi beberapa faktor lain seperti dimensi dari tes, format jawaban responden, dan jumlah sampel yang digunakan. Bahkan, keberhasilan dari IRT terletak pada prosedur yang memadai yang digunakan dalam estimasi parameter tersebut.

Estimasi parameter dapat dilakukan dalam beberapa cara. Namun yang paling banyak digunakan adalah metode maximum likelihood.

Estimasi maximum likelihood membutuhkan jumlah sampel yang cukup besar dan penggunaan estimasi ini dapat diaplikasikan dalam berbagai model. Namun sebenarnya estimasi kemampuan individu dalam IRT tidak hanya terbatas pada estimasi maximum likelihood ada beberapa prosedur estimasi lain diantaranya regresi logistik (Reynolds, Perkins & Brutten dalam Ching-Fung, 2002), *minimum chi-quadrant* (Zwinderman & van der Wollenberg dalam Ching-Fung, 2002) dan prosedur estimasi model Bayesian (Mislevy, Baker dalam Ching-Fung, 2002). Namun, penelitian kali ini menggunakan estimasi Bayesian dikarenakan terdapat beberapa situasi yang tidak dapat diselesaikan menggunakan estimasi maximum likelihood.

Dalam Hambleton (1991) fungsi Likelihood (atau log-likelihood) memiliki keterbatasan seperti (a) ketika responden menjawab semua item dengan benar atau salah, yang estimasi maximum likelihood-

nya dinyatakan $\theta = +\infty$ atau $\theta = -\infty$, (b) ketika terdapat beberapa pola respon yang aneh. Di samping itu, salah satu ciri dari estimasi maximum likelihood ialah asimptotik, dimana diperlukan sampel yang besar dan menggunakan tes yang panjang (item yang cukup banyak), sehingga *theta* (kemampuan responden) pada estimasi maximum likelihood akan terdistribusi secara normal dan tidak bias. Namun pada kenyataannya jarang ditemui bahwa estimasi terhadap kemampuan responden dilakukan dengan peserta yang jumlahnya ribuaan seperti pada seleksi pegawai atau penerimaan mahasiswa baru dan menggunakan tes dengan jumlah item yang banyak (jarang menggunakan item di atas 200). Untuk kondisi yang demikian estimasi Bayesian lebih presisi digunakan untuk mengestimasi kemampuan responden dalam jumlah yang besar dengan item yang sedikit. Penjelasan mengenai estimator Bayesian akan dipaparkan dalam metode penelitian.

Selanjutnya, sebelum mempergunakan IRT (Hambleton, 1991) hal yang terpenting harus diperhatikan ialah terpenuhinya dua asumsi dasar yakni unidimensi (*unidimensionality)* dan independensi lokal *(local independence)*. Unidimensi diartikan bahwa apa yang diukur melalui beberapa kumpulan item atau soal hanya mengukur satu *traits.* Sedangkan, asumsi *local-independence* dimaknai sebagai kemampuan individu item dalam performa tes dianggap konstan dan respon terhadap setiap item yang dijawab adalah *independent* (tidak saling bergantung). Kemampuan yang dinyatakan dalam model adalah satu-satunya faktor yang mempengaruhi respon peserta tes pada butir-butir soal.

Unidimensi dalam IRT merupakan syarat yang harus dipenuhi dimana item tersebut mendefinisikan satu konstruk utama atau dimensi. Jika ada banyak item yang tidak sejajar dengan konstruk utama, maka dapat diartikan sebagai multidimensi dan lebih dari satu. Situasi IRT yang memenuhi asumsi unidimensi atau homogenitas item jarang terjadi baik dalam bidang dipendidikan maupun psikologi. Hal ini mungkin disebabkan selain dari faktor kognitif, juga dipengaruhi oleh *personality* responden dalam menjawab item pertanyaan yakni kecepatan kerja, instruksi yang ada, *guessing* atau kecenderungan menebak. Selain dari diri responden, faktor tersebut juga bisa berasal dari rangsangan item soal yang sedang diberikan seperti panjangnya teks (pertanyaan ataupun pernyataan yang ada), tabel, gambar, peta, atau grafik yang tersaji pada soal. Sebagai contoh, tes matematika dengan item pertanyaan yang sangat panjang dan berbelit-belit akan menyebabkan responden (siswa) sulit untuk memahami isi pertanyaan dari soal tersebut, dan membutuhkan kemampuan membaca yang cukup besar. Ketika berhadapan responden dengan latar belakang yang berbeda, beberapa diantaranya mungkin cukup

mahir untuk membaca dan memahami soal cerita matematika, akibatnya kemampuan membaca mungkin sekunder dimensi (Almond, Heath, Helwig, Rozek-Tedesco & Tindal, dalam Bo Zhang, 2008). Disamping itu, adanya gambar, tabel, grafik atau peta yang ada pada soal untuk menyelesaikan pertanyaan akan menyulitkan siswa dengan kemampuan imajinasi gambar yang lemah.

Ketika asumsi unidimensional sudah terpenuhi, tahapan selanjutnya ialah bagaimana memperlakukan penskoringan unidimensional pada sebuah tes apabila terdapat beberapa dimensi yang membentuk di dalamnya. Hal yang biasa dan paling sering dilakukan ialah memperlakukannya dan menganggap sebagai unidimensional pada tingkat pertama atau *first order*. Seperti yang dilakukan oleh guru-guru di sekolah misalnya pada pelajaran bahasa inggris materi yang diujikan terdiri dari *reading, listening* dan *writing*, ataupun pelajaran matematika yang terdiri dari beberapa sub materi misalnya logartima, persamaan fungsi kuadrat, trigonometri dan ruang tiga dimensi. Sebagai nilai akhir, guru hanya memberikan satu nilai tiap pelajaran tertentu dari beberapa sub materi yang diujikan didalamnya. Tidak hanya bidang pendidikan yang memperlakukan kondisi tersebut, namun untuk bidang psikologi hal ini tampaknya juga masih banyak diterapkan. Seorang peneliti masih menskor sebuah skala yang digunakan untuk mengukur perilaku ataupun persepsi mengenai sesuatu yang terdiri dari beberapa dimensi diperlakukan sebagai satu nilai. Misalnya skala kepribadian *big* five yang terdiri dari lima dimensi yakni *neuroticism, extraversion, openness to experience, agreeableness,* dan *conscientiousness*, diperlakukan dengan menskor semua item pernyataan sebagai satu kesatuan.

Sebagian besar sistem penskoran masih memperlakukan unidimensional *first order* terhadap tes yang didalamnya terdiri dari beberapa dimensi. Lalu bagaimana hasilnya jika *scoring* atau penskoran diperlakukan dengan cara unidimen-sional *second order*. Ini diartikan bahwa sebuah tes yang terdiri dari beberapa dimensi di dalamnya, terlebih dahulu diskor pada dimensi masing-masing, kemudian nilai kesemua dimensi tersebut diestimasi untuk mendapatkan nilai kumulatif dari kesemua dimensi yang ada sehingga nilai inilah yang dianggap sebagai kemampuan respon-den pada tes tersebut. Unidimensional *second order* rasanya belum sering ditemui dilapangan. Untuk evaluasi belajar di sekolah, unidimensional *second order* sebaiknya dilakukan karena apabila guru memberikan penilaian terhadap sub-bab atau dimensi dan tidak langsung mem-berikan *final score*, maka dapat membantu siswa mengetahui dimana letak kekurangan atau ketidak-mampuan pada sub-bab tertentu.

Unidimensional *second order* ialah model pengukuran yang terdiri dari dua tingkat. Tingkat pertama

menjelaskan hubungan antara variabel observed atau variabel measured dengan variabel laten, sedangkan pada tingkat kedua menjelaskan hubungan antara variabel laten di tingkat pertama dengan variabel laten di tingkat ke dua (Joreskog dan Sorbom, 1996). Persamaan analisis faktor *second order* model y, yakni

$$Y = \lambda_y (\Gamma\xi + \zeta) + €$$

Keterangan:

λ_y = matriks faktor loading dari *first order*, dimana baris dari matrik merupakan banyaknya variabel observed dan kolom dari matriks ialah banyaknya variabel laten.

Γ = matriks faktor loading dari *second order*, dimana baris dari matrik merupakan banyaknya variabel laten pada *first order* dan kolom dari matriks ialah banyaknya variabel laten pada *second order*.

ξ = vector dari faktor variabel latent pada *second order*

ζ = vector dari komponen unik atau error pada *second order*

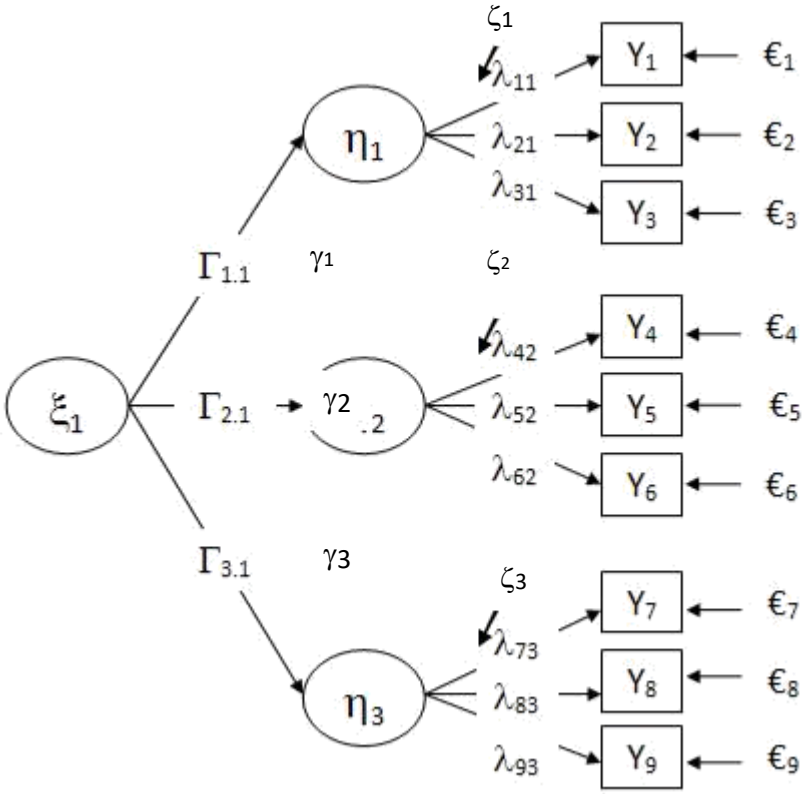
$€$ = vector dari komponen unik atau error pada *first order*

ϕ = matriks kovarians dari faktor variabel latent pada *second order*, dimana baris dan kolom pada matriks merupakan banyaknya variabel laten pada *second order*

Ψ = matriks kovarians dari komponen unik atau error pada *second order*, yang biasanya diagonal

$\Theta_€$ = matrik kovarians dari komponen unik atau error pada *first order*, dimana baris dan kolom pada matriks merupakan banyaknya error, biasanya juga diagonal

Gambar di bawah ini adalah bentuk dari analisis faktor model y dengan *first order* faktor η dan error pengukuran $€$ dengan Y sebagai variabel observednya, sehingga menjadi $Y = \lambda_y \eta + €$. Sekarang, variabel η digantikan dengan set faktor dari ξ, sehingga disebut dengan faktor *second order*, bahwa $\eta = \Gamma\xi + \zeta$. Dimana Γ adalah matrix faktor loading *second order* dan ζ adalah vector dari variabel unik untuk η. Gabungan dari $Y = \lambda_y \eta + €$ dan $\eta = \Gamma\xi + \zeta$ memberikan $Y = \lambda_y (\Gamma\xi + \zeta) + €$ dengan matrik kovarians $\sum = \lambda_y (\Gamma\phi\Gamma' + \Psi) \lambda_y' + \Theta_€$. $\sum$ inilah yang digunakan untuk menguji Ho: $S - \sum = 0$.



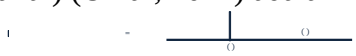
Berdasarkan uraian di muka, peneliti tertarik untuk melihat bagaimanakah sebuah tes jika memiliki model unidimensional *second order* namun biasanya diperlakukan sebagai unidimensional *first order*. Peneliti tertarik untuk menemukan jawaban dari beberapa pertanyaan, diantaranya: (1) apakah terdapat perbedaan hasil dalam mengestimasi kemampuan seseorang jika data yang ada memiliki model *second order unidimensional*, namun diperlakukan sebagai *first order unidimensional*; (2) dalam kondisi seperti apakah *second order unidimensional* bisa diperlakukan sama seperti *first order unidimensional;* (3) bagaimanakah pengaruh panjang tes, heterogenitas tingkat kesukaran dan daya pembeda jika terdapat bias antara *second order unidimensional* dan *first order unidimensional*.

Metode Penelitian

Penelitian ini merupakan penelitian simulasi Monte Carlo yang dirancang dan dibayangkan apabila terjadi di dunia nyata atau sebenarnya. *Software* yang digunakan untuk membangkitkan data serta menganalisisnya sebagai *first order* dan *second order* ialah MPlus dengan

bantuan estimator Bayesian (Muthen, 2010). Ide dasar estimator Bayesian adalah memodifikasi fungsi likelihood dengan cara memasukkan informasi sebelum kita mendapatkan parameter kemampuan. Prosedur bayesian merupakan prosedur yang menggunakan atau menggabungkan pengetahuan subjektif (terdahulu) tentang para-meter yang akan ditaksir dengan informasi yang diperoleh dari data sampel. Informasi terdahulu disebut disebut juga dengan informasi prior, diperoleh dari distribusi parameter
tersebut. Informasi dari data dirangkum dalam fungsi likelihood. Penggabungan dari informasi prior dan informasi dari data akan menghasilkan informasi posterior.

Teorema Bayes menyatakan bahwa probabilitas kondisional (*conditional probability)* dari suatu peristiwa A jika peristiwa B sudah
terjadi (probabilitas terjadinya peristiwa A jika kondisi B sudah diketahui) (Umar, 2012) adalah



Di dalam pendekatan bayesian, estimasi parameter pada sebuah model statistik, dilakukan dengan cara menyederhankan rumusan di atas menjadi persamaan yang bersifat proporsionalnya yakni (Hambleton, 1991):

$$P(A \mid B) \propto P(B \mid A)\, P(A)$$

dimana simbol $\propto$ diartikan sebagai "proporsional terhadap", A sebagai hipotesis atau parameter sedangkan B merupakan data yang diperoleh.

Rumusan di atas diperoleh karena P(B) dalam perhitungannya bersifat konstan. Di dalam rumus tersebut probabilitas P(A|B) *(posterior)* adalah sama dengan likelihood dari data B dalam kondisi berlakunya hipotesis A (P(B|A)) dikalikan (diboboti) dengan probabilitas hipotesis A *(prior)*. Dengan kata lain probabilitas benar tidaknya hipotesis A dalam kondisi data B sudah diperoleh (P(A|B)), adalah sama dengan probabilitas dari data dalam kondisi hipotesis A berlaku (likelihood) dikalikan dengan probabilitas hipotesis A yang berdasarkan pengalaman atau pengetahuan yang sudah ada sebelumnya (prior). Pada konteks ini P(A|B) disebut posterior. Sebagai kesimpulan atau ringkasnya, pobabilitas posterior adalah likelihood yang dikoreksi atau disesuaikan dengan probabilitas prior (pengetahuan atau teori yang telah dimiliki sebelumnya).

Rumusan diatas dapat juga ditulis dalam bentuk:

$$\text{posterior} \propto \text{likelihood} * \text{prior}$$

Ini diartikan bahwa likelihood dari
data digunakan sebagai bahan untuk memperbaharui informasi prior sehingga menjadi sebuah informasi posterior yang siap dipakai. Hubungan di atas juga berlaku untuk fungsi padat (densitas), dimana A adalah theta (θ) dan B adalah pola respon item yang teramati (u). Dalam Hambleton (1991) teorema bayes dapat ditulis seperti dibawah ini

$$f(\theta \mid u) \propto f(u \mid \theta)\, f(\theta)$$

di mana $f(\theta)$ adalah distribusi prior dari hipotesis atau pengetahuan. Selanjutnya, karena $f(u \mid \theta)$, pada kenyataannya adalah fungsi likelihood maka persamaan di atas dapat ditulis sebagai:

$$f(\theta \mid u) \propto L(u \mid \theta) f(\theta)$$

Setelah menentukan software dengan estimator yang akan digunakan maka penelitian simulasi ini melakukan beberapa langkah, yakni.

Pertama, sesuai dengan tujuan penelitian simulasi, maka dibutuhkan desain penelitian yang nantinya akan mempermudah membangkitkan data serta menganalsisnya. *Independent variable* dalam penelitian ini ialah panjang tes, heterogenitas tingkat kesukaran dan daya pembeda. Panjang test yang disimulasi mewakili test pendek dan tes panjang. Sesuai dengan pernyataan Mislevy & Bock (1990), tes pendek merupakan tes yang terdiri dari 11 sampai 20 soal, sedangkan tes panjang lebih dari 20 soal. Oleh sebab itu, dalam penelitian ini menggunakan 20, 40 dan 60. Tes dengan panjang 20 mewakili tes pendek sedangkan tes dengan panjang 40 dan 60 mewakili tes panjang. Heterogenitas tingkat kesukaran dan daya pembeda ditentukan dengan nilai varians 0,025, 0,10 dan 0,20. Berdasarkan panjang tes, heterogenitas tingkat kesukaran dan daya pembeda akan ada 3 x 3 x 3 = 27 model data yang dibangkitkan seperti tabel di bawah ini:

Skema Simulasi Sebanyak 27 Model

TS / DB	Panjang Tes								
	20			40			60		
	0,025	0,10	0,20	0,025	0,10	0,20	0,025	0,10	0,20
0,025	AA	AB	AC	AA	AB	AC	AA	AB	AC
0,10	BA	BB	BC	BA	BB	BC	BA	BB	BC
0,20	CA	CB	CC	CA	CB	CC	CA	CB	CC

Keterangan:
Panjang tes terdiri dari 20, 40 dan 60 item
DB = Daya beda dengan heterogenitas 0,025, 0,10 dan 0,20
TK = Tingkat kesukaran dengan heterogenitas 0,025, 0,10 dan 0,20

Keseluruhan model akan direplikasi sebanyak 50 kali yang nantinya akan dianalisis sebagai *second order* dan *first order* unidimensional data. Penentuan jumlah replikasi ini mengacu pada penelitian sebelumnya yang dilakukan oleh Kamata (dalam Ching-Fung, 2002), ia melakukan replikasi sebanyak 5 hingga 50 dan menemukan bahwa estimasi parameter kemampuan dan item akan stabil setelah direplikasi lebih dari 50 kali. Sementara itu, *dependent variabel* ialah jumlah dimensi, nilai muatan

faktor loading gamma (Γ) yakni dari eta (η) ke ksai (ξ), dan jumlah responden yang sama pada setiap model yakni empat dimensi dengan 1000 responden atau *examinee*.

Kedua, untuk memastikan data yang sudah dibangkitkan memiliki model *second order unidimensional* dilakukan uji *exploratory factor analysis* (EFA) dan *confirmatory factor analysis* (CFA) pada beberapa replikasi di keseluruhan model. Hasil analisis data menggunakan EFA dengan cara melihat *eigenvalue* di atas satu harus sebanyak empat buah. Jika hal ini terjadi maka data yang berhasil dibangkitkan memiliki empat dimensi pada *second* order unidimensional. Selanjutnya, untuk mengetahui apa-kah keempat dimensi tersebut membentuk *second order,* maka dilakukan uji CFA. Hasil dari CFA yang membentuk *second order* ditunjukan P-Value di atas 0.05 (tidak signifikan). Ini artinya tidak ada perbedaan antara data replikasi dengan model yang ada atau diinginkan. Namun, apabila P-Value lebih kecil dari 0.05 maka data replikasi yang dibangkitkan tidak membentuk *second* order unidimen-sional.

Ketiga, data *second order unidimensional* yang sudah dibangkit-kan akan dianalisis sebagai *first order unidimensional* dan *second order unidimensional*. Nilai *first* order dianggap sebagai *estimate* sedangkan nilai *second order* dianggap sebagai *true score.* Untuk melihat apakah terdapat perbedaan atau bias antara *first order* dan *second order* pada data *second order unidimensional,* maka peneliti menghitung nilai selisih atau bias atau error atau deviasi dari analisis tersebut, dengan rumus:

$$\text{Bias atau Deviasi} = \hat{\theta} - \theta$$

Keterangan:

$\hat{\theta}$ = theta estimate (hasil analisis menggunakan first order unidimensional)
θ = theta true (hasil analisis menggunakan second order unidimensional)

Hasil analisis pada *second order unidimensional data* namun diper-lakukan sebagai *first order unidimen-sional* memiliki atau menghasilkan bias atau deviasi jika hasil pengurangan *theta estimate* dengan *theta true* lebih besar dari nol. Namun jika hasil pengurangan kedua *theta* sama dengan nol maka tidak terdapat bias atau deviasi pada data *second order unidimensional* tetapi sering diperlakukan sebagai *first order unidimensional.*

Hasil Penelitian

Di dalam setiap model terdiri dari 50 replikasi, dimana tiap replikasinya akan memiliki mean dan standar deviasi dari 1000 bias responden (dengan menggunakan rumus di atas). Maka sebuah model akan memiliki 50 nilai mean dan

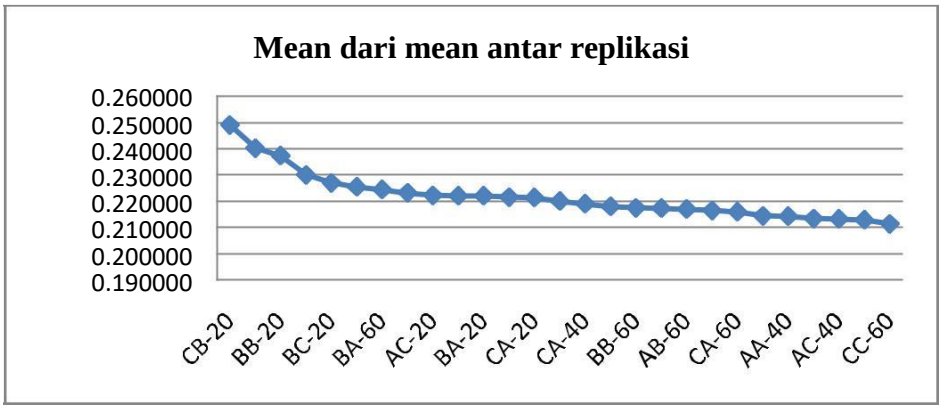
standar deviasi. Dari 50 nilai mean tersebut (antar replikasi) akan diperoleh nilai mean dan standar deviasi. Nilai ini yang digunakan untuk melihat rata-rata bias yang dihasilkan antar replikasi dalam sebuah model. Berikut hasil perhitungan mean dan standar deviasi dari mean bias responden antar replikasi untuk keseluruhan model:

Mean dari Mean Bias Responden Antar Replikasi

TK \	Panjang Tes								
	20			40			60		
DB	0.025	0.10	0.20	0.025	0.10	0.20	0.025	0.10	0.20
0.025	0.217970	0.221603	0.222158	0.214270	0.213413	0.213201	0.216436	0.216982	0.212910
0.10	0.221981	0.237308	0.226984	0.217351	0.222092	0.223161	0.224492	0.217380	0.214382
0.20	0.221388	0.249048	0.225523	0.219023	0.230012	0.240291	0.215902	0.220012	0.211427

Keterangan:
Panjang tes terdiri dari 20, 40 dan 60 item
DB = Daya beda dengan heterogenitas 0.025, 0.10 dan 0.20
TK = Tingkat kesukaran dengan heterogenitas 0.025, 0.10 dan 0.20

Hasil nilai mean di atas memang tidak dibulatkan dua angka dibelakang koma dikarenakan perbedaan mean yang dihasilkan oleh setiap model nantinya tidak terlihat jelas. Pada tabel di atas dapat dilihat nilai mean bias antar replikasi untuk 27 model, lebih besar dari nol artinya terdapat perbedaan antara analisis *first* order dengan *second* order pada data *second* order unidimensional. Perbedaan diantara 27 model tidak terlalu bervariasi karena nilai yang dihasilkan berkisar antara 0.211427 hingga 0.249048. Untuk lebih jelasnya berikut grafik mean dari mean bias responden antar replikasi:



Dari grafik di atas, dapat dilihat bahwa puncak tertinggi hasil penghitungan mean dari mean bias responden antar replikasi terdapat pada panjang tes 20 item dengan model CB (daya beda 0.20 dan tingkat kesukaran 0.025), sedangkan titik terendah dari grafik tersebut terdapat pada tes dengan panjang 60 item model CC (daya beda dan tingkat
kesukaran 0.20). Namun apabila dilihat secara kasat mata nilai mean dari mean bias responden antar replikasi untuk semua model perbedaannnya tidak terlalu jauh. Hal ini dapat dilihat penurunan grafik tidak terlalu curam dan nilai mean pada grafik berkisar antara 0.25 hingga 0.21.

Dari nilai mean di atas, dapat dikelompokan menjadi mean tinggi, mean sedang dan mean rendah, seperti dapat dilihat pada tabel di bawah ini:

TK / DB	Panjang Tes								
	20			40			60		
	0.025	0.10	0.20	0.025	0.10	0.20	0.025	0.10	0.20
0.025									
0.10									
0.20		Tinggi							Rendah

Keterangan:

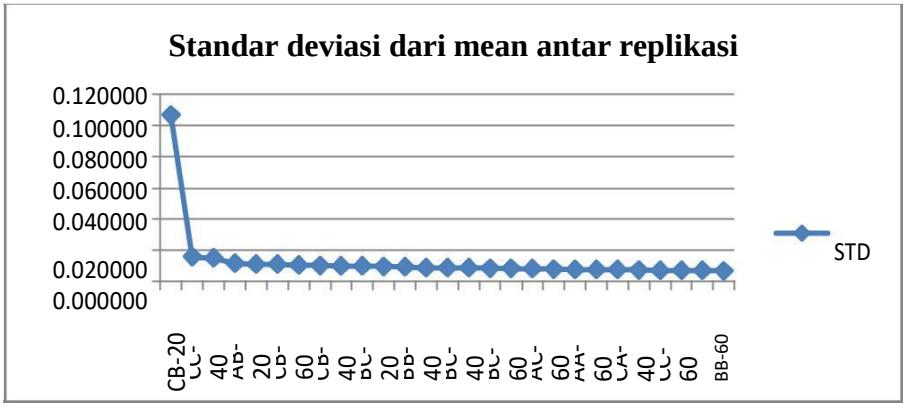
Pada tabel di atas, mean di kelompok tinggi didominasi oleh tes dengan panjang 20 item dan 40 item, hanya satu model dengan panjang tes 60 item yang termasuk di dalamnya. Dilihat dari daya beda dan tingkat kesukarannya dalam mean tinggi tidak ada model dengan heterogenitas AA (daya beda 0.025 & tingkat kesukaran 0.025), AB (daya beda 0.025 & tingkat kesukaran 0.10) dan CA (daya beda 0.20 & tingkat kesukaran 0.025). Untuk mean di kelompok sedang, masih didominasi oleh model dengan panjang tes 20 item, panjang tes 40 item dan panjang tes 60 item sebanyak dua model. Sedangkan untuk heterogenitas daya beda dan tingkat kesukaran yang tidak ada dalam mean kelompok sedang ialah model AC, BC dan CC (daya pembeda 0.025, 0.10 dan 0.20 dengan tingkat kesukaran yang sama 0.20). Untuk mean pada kelompok rendah sangat didominasi dengan panjang tes 60 item sebanyak enam model, tiga model dengan jumlah item 40 dan tes dengan panjang 20 item tidak ada satupun di dalam kelompok mean rendah. Heterogenitas daya pembeda dan tingkat kesukaran yang tidak ada di kelompok tinggi yakni BA (daya

beda 0.10 & tingkat kesukaran 0.025), BB (daya beda & tingkat kesukaran 0.10) dan CB (daya beda 0.20 & tingkat 0.10).

Selanjutnya untuk melihat bagaimana fluktuasi atau bervariasi-nya mean bias responden antar replikasi satu dengan yang lainnya dalam setiap model yang ada, maka dihitung standar deviasi dari mean bias responden antar replikasi untuk keseluruhan model, seperti table di bawah ini:

TK / DB	Panjang Tes								
	20			40			60		
	0.025	0.10	0.20	0.025	0.10	0.20	0.025	0.10	0.20
0.025	0.009926	0.011219	0.011182	0.007313	0.007838	0.008194	0.007809	0.008884	0.007998
0.10	0.010318	0.015945	0.009668	0.007102	0.008980	0.008862	0.009578	0.007032	0.008538
0.20	0.008626	0.106832	0.011850	0.007547	0.009967	0.015140	0.007750	0.010622	0.007175

Dari tabel di atas dapat dilihat bahwa standar deviasi dari mean bias responden antar replikasi yang dihasilkan sangat kecil bahkan mendekati nol. Ini artinya mean bias antar replikasi satu dengan yang lainnya sebanyak 50 kali replikasi dalam setiap model tidak bervariasi atau homogen, apabila diteruskan untuk mereplikasi maka nilai mean bias yang dihasilkan akan sama. Namun jika dilihat pada tabel nilai standar deviasi dari mean bias responden antar replikasi paling besar dengan nilai 0.106832 ada pada model 20-CB yakni panjang tes 20 item, daya beda 0.20 dan tingkat kesukaran 0.10. Sedangkan nilai standar deviasi dari mean bias responden antar replikasi paling kecil dengan nilai 0.007032 dihasilkan pada panjang tes 60 item dengan daya beda dan tingkat kesukaran 0.10. Untuk lebih jelas melihat bagaimana perbedaan standar deviasi dari mean bias responden antar replikasi dalam setiap model, maka peneliti sertakan grafik dari nilai tersebut di bawah ini:



Pada grafik jelas terlihat panjang tes 20 item dengan model CB (daya beda 0.20 dan tingkat kesukaran 0.10) memiliki puncak yang paling tinggi. Sedangkan titik terendah dari grafik dimiliki oleh model 60-BB yakni panjang tes 60 item, daya beda dan tingkat kesukaran 0.10. Penurunan grafik terlihat jelas dari model 20-CB ke model 20-BB dari sekitar nilai 0.10 ke 0.02, namun setelah itu grafik terlihat konstan dengan nilai antara 0.20 hingga mendekati 0.00.

Nilai mean bias responden antar replikasi dalam setiap model dapat digunakan untuk mengetahui sejauh apa interaksi dari pengaruh heterogenitas daya pembeda, tingkat kesukaran dan panjang tes serta pengaruh heterogenitas tingkat kesukaran dan panjang tes jika daya pembeda dalam nilai yang sama. Berikut hasil penghitungannya:

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected model	.102[a]	26	.004	7.635	.000
Intercept	66.372	1	66.372	129380.222	.000
Item	.025	2	.012	24.005	.000
Dayabeda	.020	2	.010	19.703	.000
Kesukaran	.010	2	.005	9.683	.000
Item * Dayabeda	.011	4	.003	5.254	.000
Item * Kesukaran	.018	4	.005	8.858	.000
Dayabeda * Kesukaran	.007	4	.002	3.514	.007
Item * Dayabeda * Kesukaran	.011	8	.001	2.654	.007
Error	.679	1323	.001		
Total	67.152	1350			
Corrected Total	.781	1349			

a. R Squared = .130 (Adjusted R Squared = .113)

Dari tabel di atas dapat diketahui bahwa interaksi antara panjang tes, heterogenitas daya beda dan tingkat kesukaran memiliki nilai R^2 = 0.130 dan nilai signifikansi = 0.007. Maka dapat dikatakan bahwa pengaruh panjang tes, heterogenitas daya beda dan tingkat kesukaran bisa meramalkan 13% dari mean bias atau perbedaan antara *first* order dengan *second* order pada data *second* order unidimensional.

Kesimpulan

Penelitian yang dilakukan terhadap *second order unidimensional data* namun sering diperlakukan sebagai *first order unidimensional* menghasilkan kesimpulan: (1) terdapat bias atau perbedaan antara

keduanya sehingga *theta* (θ atau kemampuan responden) yang diperoleh atau dihasilkan melalui analisis *first order unidimensional* tidak menggambarkan keadaan yang sebenarnya; (2) setiap replikasi dari keseluruhan model percobaan menghasilkan nilai lebih besar dari nol artinya setiap replikasi yang dilakukan menunjukan bias antara *first order* dengan *second order*; (3) dari 27 model percobaan, secara rata-rata mean bias paling besar dihasilkan panjang tes yang paling kecil yakni 20 item dengan tingkat kesukaran 0,10 dan daya beda 0,20, sedangkan mean bias paling kecil dihasilkan oleh panjang tes paling besar yakni 60 item dengan tingkat kesukaran dan daya pembeda 0,20; (4) dilihat dari kelompok mean tinggi, sedang dan rendah, secara rata-rata maka tes dengan panjang 20 item akan menghasilkan mean bias tinggi dari keseluruhan model, mean bias sedang dihasilkan oleh tes dengan panjang 40 item dan tes dengan panjang 60 item menghasilkan mean bias rendah; (5) terdapat interaksi yang signifikan antara panjang tes, heterogenitas tingkat kesukaran dan daya beda sebesar 13% dari mean bias atau perbedaan antara *first* order dengan *second* order pada data *second* order unidimensional.

Bagi para peneliti atau mahasiswa yang memiliki data unidimensional pada *second* order namun menganalisisnya hanya pada *first* order unidimensional alangkah baiknya hal ini tidak dilakukan karena terdapat bias sehingga hasil yang ada tidak memberikan gambaran seutuhnya mengenai kemampuan responden yang sedang dianalisis. Kalaupun tetap memberlakukan *first* order pada data unidimensional *second order* bias paling kecil diperoleh dengan panjang tes yang lebih besar. Dalam penelitian ini tes dengan panjang 60 item bias yang dihasilkan lebih rendah dibandingkan tes dengan panjang 20 item.

Rekomendasi untuk penelitian berikutnya yang tertarik mengadakan penelitian dengan studi simulasi Monte Carlo menggunakan model *second order unidimensional* dapat melihat pengaruh dari variabel-variabel lain, seperti bagaimanakah pengaruh jumlah responden, pengaruh jenis distribusi, pengaruh jumlah dimensi atau faktor, atau pengaruh tinggi rendahnya nilai faktor loading gamma (Γ).

Daftar Pusataka

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Florida: Harcourt Brace Jovanich Collage Publish.

du Toit, ME. (2003). *IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact*. Lincolnwood, IL: Scientific Software International.

Embretson, S.E., & Reise, S. P. (2000). *Item response theory for psychology*. London: Lawrence Erlbaum Associates, Publishers.

Fung, C. (2002). *Ability Estimation Under Different Item Parameterization And Scoring Models*. Dissertation, University of North Texas.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory, principle and application.* Boston: Kluwer Nijhoff Publishing.

Hambleton, R.K., Swaminathan, H., & Rogers, J.H. (1991). *Fundamentals of item response theory.* California: SAGE Publications.

Joreskog, K.G., & Sorbom, Dag. (1996). *Lisrel 8, User's Reference Guide*. Chicago: SSI, Inc (Scientific Software International).

Kaplan, R.M., & Saccuzo, D.P. (1993). *Psychological Testing: Principles, Applications, and Issues.Third edition.*California: Brooks/Cole Publishing.

Mislevy, R.J. & Bock, R.D. (1990). *BILOG 3: Item analysis & test scoring with binary logistic models.* Moorseville: Scientific Software, Inc.

Muthen, L.K., & Muthen, B.O. (2010). *Mplus, statistical analysis with latent variables user's guide*. Los Angeles: StatModel.

Umar, J. (2012). *Bahan kuliah psikometri*: Analisis Faktor. Jakarta: Tidak dipublikasikan.

Umar, J. (2012). Mengenal lebih dekat konsep reliabilitas skor tes. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia.* 1: No. 2.

Zhang, B. (2008). Application of Unidimensional Item Response Models to Test With Item s Sensitive to Secondary Dimension. *The Journal of Experimental Education*. 77(2), 147-166