DAMPAK DARI MULTIDIMENSIONALITAS BUTIR SOAL TERHADAP ESTIMASI *TRUE SCORE* DENGAN PENDEKATAN MODEL BIFAKTOR

Nursakinah Oktaviana Sasmita
UIN Syarif Hidayatullah Jakarta
oktaviana_sakinah@gmail.com

Abstract

Current study is a simulation research, focused on the number of factors, items*, and respondents replicated 50 times. After that, replicated data was considered as unidimensional and bifactor and then the effect was computed from theta margin. This research aims to explore the number of factors, items, and respondents, which is measured, affect the unidimensional asumption transgression on bifactor. Also, this research aims to understand bias differences of bifactor data that is considered as unidimensional. The result showed that data with bifactor model and analyzed as unidimensional will obtain the untrue theta score due to high bias differences. In addition, the R square of respondents bias is 0.69%.*

***Keywords**: Bifactor Model, Unidimensional*

Abstrak

Penelitian ini merupakan penelitian simulasi dimana yang menjadi fokus dalam penelitian ini adalah banyaknya faktor, item dan responden dengan replikasi 50 kali. Selanjutnya data hasil replikasi ini dianggap sebagai unidimensi dan bifaktor dan dihitung pengaruhnya dari selisih theta tersebut. Penelitian ini bertujuan untuk dapat mengetahui banyaknya faktor, item dan responden yang ikut terukur berdampak pelanggaran asumsi unidimensi pada bifaktor. Selain itu, juga untuk mengetahui perbedaan bias pada data bifaktor yang dianggap sebagai unidimensi tersebut. Hasil penelitian menunjukkan bahwa data dengan model bifaktor dan dianalisis sebagai unidimensi maka hasilnya akan memperoleh theta yang tidak sebenarnya, karena perbedaan bias atau deviasi yang terjadi cukup tinggi. Disamping berdasarkan hasil perhitungan didapatkan R square *sebesar 0.69%, bias responden yang dapat dijelaskan oleh bervariasinya faktor,* item *dan responden dengan taraf signifikansi 0.000.*

***Kata Kunci**: Model Bifaktor, Unidimensi*

Diterima: 3 April 2015 Direvisi: 27 April 2015 Disetujui: 10 Mei 2015

PENDAHULUAN

Tes merupakan salah satu alat pengukuran yang paling sering digunakan pada bidang pendidikan dan psikologi. Pada pelaksanaannya, tes seharusnya berasaskan objektif, transparan, akuntabel dan tidak diskriminatif. Suatu alat tes, sebaiknya hanya bersifat unidimensi yang artinya setiap item tes hanya mengukur satu kemampuan. Asumsi hanya dapat ditunjukkan jika tes mengandung satu faktor yang mengukur prestasi suatu subjek.

Tujuan penggunaan tes biasanya banyak sekali ragamnya, namun selalu berkenaan dengan satu hal, yaitu penggunaan skor tes untuk mengambil suatu keputusan. Metode-metode psikometri dapat dipergunakan sebagai alat untuk mencapai dua jenis tujuan, diantaranya untuk mendapatkan persamaan matematis yang paling handal dalam meramalkan akibat dari suatu keputusan yang akan diambil berdasarkan skor dari satu atau sehimpunan tes, dan untuk menguji apakah suatu model teoritis tentang cara penggunaan skor tes tertentu untuk tujuan tertentu yang selama ini mungkin telah sering dipakai yang memang cukup handal dan dipercaya (Umar, 2011).

Pengukuran yang menggunakan alat ukur yang baik, maka akan baik pula data yang diperoleh sehingga memudahkan dalam evaluasi dan interpretasi data tersebut. Apabila di dalam suatu pengukuran alat tes tersebut terdapat kesalahan dalam metodologi pengukurannya maka akan berdampak fatal terhadap nilai tes dan akan merugikan berbagai macam pihak. Khususnya yang mengikuti tes tersebut. Apalagi jika tes tersebut merupakan tes intelegensi untuk masuk ke dalam suatu perusahaan. Responden yang nilainya rendah bukan dikarenakan intelegensinya rendah, namun bisa disebabkan berbagai macam sebab. Hal ini berdampak, kehilangan calon karyawan yang sebenarnya mempunyai kualitas yang tinggi tetapi tidak lulus karena nilainya yang tidak mencapai persyaratan atau lebih parahnya lagi karena metode yang digunakan dalam mengukur alat tes tersebut salah.

Kesalahan metode yang digunakan dalam menganalisis hasil tes, dapat disebabkan karena kurangnya pengetahuan tentang metode yang seharusnya digunakan dalam menganalisis hasil tes tersebut. Umumnya, analis hasil tes menggunakan teori tes klasik yang telah mendominasi dan banyak berjasa di bidang pengukuran. Esensi dari teori tes klasik berupa asumsi-asumsi yang dirumuskan secara matemastis. Tetapi seiring berjalannya waktu, teori tes klasik memiliki kelemahan-kelemahan. Kelemahan utama dalam teori ini adalah bahwa alat ukur yang disusun berdasarkan teori tersebut memiliki keterikatan terhadap sampel yang digunakan. Padahal suatu tes dianggap baik apabila tes tersebut memiliki sifat, dalam arti tes tersebut tidak terikat (bebas) dari jenis sampel yang digunakan, misalnya alat-alat ukur fisik (alat ukur panjang dan berat) yang kesemuanya tidak terikat pada sampel yang digunakan. Permasalahan dalam faktor pengukuran psikologis lainnya adalah tidak ada pendekatan tunggal dalam pengukuran (perbedaan teori dapat menyebabkan pula perbedaan objek ukur), perilaku manusia tidak terbatas (permasalahan pengambilan sampel perilaku), adanya unsur eror dalam pengukuran (permasalahan konsistensi dan ketepatan pengukuran), satuan dalam pengukuran (permasalahan interpretasi hasil pengukuran), dan hubungan dengan konstrak lain (hasil pengukuran dikaitkan dengan fenomena lain yang dapat diamati) (Widhiarso, 2011).

Keterbatasan pada teori tes klasik tersebut diungkap Hambleton, Swaminathan & Rogers (1991), yakni adanya sifat *group dependent* dan item *dependent*, juga indeks daya pembeda, koefisien validitas, koefisien reliabilitas skor tes yang keseluruhannya tergantung kepada peserta tes yang mengerjakan tes tersebut.

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

Item *dependent* artinya hasil pengukuran tergantung pada tes mana yang diujikan. Jika tes yang diujikan mempunyai tingkat kesulitan tinggi, estimasi kemampuan peserta tes akan rendah. Sebaliknya, jika tes yang diujikan mempunyai tingkat kesulitan rendah, estimasi kemampuan peserta tes akan tinggi.

Selain teori tes klasik, ada pula teori tes modern atau yang sering kita dengar IRT (Item *Response Theory).* Teori ini mengklaim bisa membebaskan dari keterikatan terhadap sampel. Hal ini disebabkan teori ini mendasarkan pada item, bukan lagi pada perangkat tes. Teori ini mempunyai orientasi pada item yang karakteristiknya tidak tergantung pada kelompok tertentu.

Dalam penggunaan IRT, harus memenuhi dua asumsi dasar yakni, unidimensional dan independensi lokal. Unidimensi diartikan bahwa apa yang diukur hanya mengukur satu trait. Asumsi ini sangat sulit untuk dipenuhi, karena banyaknya faktor lain yang dapat mempengaruhi seperti, motivasi, kecemasan dan lain sebagainya. Sedangkan asumsi independensi lokal diartikan sebagai kemampuan individu item dalam performa tes dianggap konstan dan respon terhadap setiap item yang dijawab adalah tidak saling bergantung.

Ada tiga model IRT (Hambleton, dkk.1991), yaitu: (1) model satu parameter (Model Rasch), yaitu hanya menitikberatkan pada parameter tingkat kesukaran item; (2) model dua parameter, yaitu hanya menitikberatkan pada parameter tingkat kesukaran dan daya pembeda item; (3) model tiga parameter, yaitu hanya menitikberatkan pada parameter tingkat kesukaran item, daya pembeda item, dan *pseudo guessing*.

Menurut Hambleton (1991), keunggulan yang dimiliki IRT (pola jawaban responden) 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 realibitas dan model menyediakan ukuran yang tepat untuk setiap skor kemampuan.

Untuk mengestimasi kemampuan responden, dalam teori respon item menggunakan data dikotomi (misal benar-salah) maupun politomi (lebih dari dua pengkategorian, misal essay atau skala likert). Data dikotomi menggunakan model matematika 1, 2 dan 3 parameter logistik. Dalam IRT, kemampuan responden dapat diperoleh dengan cara mengestimasi karakteristik parameter sesuai dengan IRT yang sedang digunakan. Penggunaan model dan parameter item yang berbeda, akan menghasilkan kemampuan yang berbeda. Dalam IRT, tidak hanya parameter item yang akan mempengaruhi hasil estimasi peserta tes, tetapi beberapa faktor lain seperti dimensi tes, format jawaban responden dan jumlah sampel yang digunakan (Lord & Novick dalam Ching-Fung, 2012).

Sebelum menerapkan IRT, asumsi pertama harus dipenuhi terlebih dahulu adalah; item tersebut harus unidimensi, artinya memiliki satu konstruk utama atau satu dimensi. Jika ada banyak item yang tidak sama dengan konstruk utama, maka item tersebut diartikan bersifat multidimensi. Situasi tersebut memang agak sulit dalam kondisi di bidang pendidikan maupun psikologi. Dalam bidang pendidikan, contohnya adalah pelajaran matematika, selain dari diri responden, faktor *guessing* (menebak) dapat juga terjadi. Kemungkinan, siswa yang mahir dalam membaca makna dari soal matematika akan dengan mudahnya menjawab soal tersebut, tetapi hal itu akan dirasa sulit bagi yang lemah dalam imajinasi gambar matematika. Banyak peneliti telah menunjukkan bahwa asumsi ini mungkin tidak berlaku untuk baterai tes tertentu, karenanya hasil dari penerapan model unidimensional ke multidimensional data tersebut dapat dipertanyakan (Kroopnick, 2010).

Sehingga jika digabungkan dengan teori respon item, yang mengasumsikan bahwa setiap item dianggap unidimensi, maka dapat diasumsikan sebagai pelanggaran terhadap teori tes. Pelanggaran dalam hal ini disebut bias. Penelitian ini penting untuk dilakukan, dengan melakukan beberapa kali simulasi pada ketiga parameter logistik. Sejalan dengan penelitian Sersano (2010), hasil dari penelitiannya dari model tanpa eror lebih buruk dari

perkiraan informasi yang lengkap berbeda dengan model kesalahan yang berkorelasi.

Dalam penelitian Simon (2008), isu bias dan konvergensi adalah salah satu masalah yang paling besar, ketika suatu kesulitan memiliki keterbatasan. Dan menurut penelitiannya, ukuran responden tampaknya memainkan peran dalam bias dan RMSE. Simon menunjukkan bahwa kalibrasi bersamaan pada umumnya lebih baik daripada metode terpisah bahkan ketika kelompok-kelompok non-setara dengan 0,5 standar deviasi perbedaan antara kelompok sarana dan korelasi antara dimensi kemampuan tinggi. Kalibrasi bersamaan memiliki manfaat yang lebih besar dari ukuran responden daripada metode menghubungkan terpisah terhadap semua parameter item, terutama dalam bentuk tes yang lebih pendek.

Penelitian ini bertujuan untuk mengetahui bagaimana dampak banyaknya faktor, banyaknya item, dan banyaknya responden terhadap estimasi parameter item dan *true score* responden jika suatu asumsi unidimensional dalam IRT dilanggar.

Dalam penelitian ini, peneliti menggunakan metode bifaktor yang jarang sekali ditemukan di lapangan. Sebagian besar sistem penskoran masih memperlakukan skor sebagai unidimensi terhadap tes yang sebenarnya memiliki beberapa dimensi. Metode bifaktor adalah bentuk dari konfirmatori faktor analisis yang dikenalkan oleh Holzinger (Jenrich dan Bentler, 2011). Metode bifaktor mempunyai faktor umum dan beberapa grup faktor. Model bifaktor digunakan apabila terdapat faktor utama yang dinyatakan sebagai penyebab bervariasinya interkorelasi antar item, dan terdapat beberapa dimensi yang menjadi penyebab interkorelasi antar *error* pada item, dimensi dan faktor utama tersebut (Chen dkk, 2006). Beberapa referensi, termasuk Chen dkk (2006), Pomplun and Simms, dkk dalam Jenrich & Bentler (2011) menerangkan bahwa model bifaktor menjadi sangat penting di lapangan Item *Responses Theory*, dimana grup faktor menjelaskan asal mula dari unidimensionalitas.

Dalam hal ini, bifaktor yang digunakan untuk dapat mengetahui pengaruh banyaknya faktor, item, dan responden yang ikut terukur berdampak pelanggaran asumsi unidimensi pada metode bifaktor, diperlukan data simulasi dengan karakteristik yang diinginkan. Penelitian ini menggunakan data simulasi, karena untuk mengetahui pengaruh tersebut, penyelesaian secara analitis tidak dapat dilakukan. Agar kondisi pada penelitian simulasi hampir sama seperti pada kondisi di dunia nyata, pada penelitian simulasi diperlukan data yang digunakan sebagai model. Dari data model ini, parameternya digunakan untuk membangkitkan data. Melalui metode pengukuran ini diharapkan dapat menggali seberapa banyak pengaruh dari banyaknya faktor, item dan responden dalam pelanggaran asumsi unidimensi pada metode bifaktor.

METODE

Simulasi adalah bidang yang berkembang pesat. Hal itu ditunjukkan dalam indeks 2000, ada 77 artikel dengan judul simulasi. Artikel tersebut berada dalam 55 jurnal yang berbeda-beda. Artikel tentang simulasi ini datang dari hampir semua disiplin ilmu sosial (Axelrod, 2005). Hal ini menunjukkan penelitian tentang simulasi penting untuk dilakukan. Tujuan dari penelitian simulasi ada bermacam-macam, diantaranya: prediksi, kinerja, pelatihan, pendidikan, bukti dan penemuan.

Simulasi dalam penelitian ini digunakan untuk memprediksi dan membantu memvalidasi suatu bukti dan penemuan. Dalam simulasi ini akan menjawab permasalahan penelitian yakni mengetahui pengaruh banyaknya faktor, panjang tes, banyaknya responden, terhadap estimasi *true score* dengan pendekatan model bifaktor. Untuk mendapatkan jawaban dari pertanyaan di atas maka dilakukan studi simulasi dengan kondisi yang akan ditentukan terlebih dahulu.

Desain Penelitian

Sesuai dengan tujuan penelitian ini, maka dibutuhkan desain penelitian untuk mengetahui pengaruh dari banyaknya faktor, item, dan responden terhadap estimasi *true score*. Data yang sudah ada lalu akan dianalisis terlebih dahulu untuk mengetahui apakah data tersebut sudah dapat dibenarkan dan sesuai dengan desain yang diinginkan peneliti.

Panjang tes yang disimulasi mewakili tes pendek, tes sedang dan tes panjang. Sesuai dengan pernyataan Mislevy dan Bock (1990), tes pendek adalah tes yang terdiri kurang dari 20 item, sedangkan tes panjang terdiri dari lebih dari 20 item. Oleh sebab itu, dalam penelitian ini akan menggunakan 10, 30 dan 80 item. Tes dengan panjang 10 item mewakili tes pendek sedangkan tes dengan item 80 mewakili tes panjang, sedangkan sebagai pelengkap, item dengan jumlah 30 adalah untuk mewakili tes sedang. Daya beda dinilai konstan 1, dan kesukaran item dinilai 1,5. Sedangkan, untuk jumlah faktor (selain faktor umum) dibagi menjadi tiga, yaitu dua, tiga, dan empat.

Untuk jumlah responden, peneliti menggunakan 250, 500 dan 1000 responden. Hal ini seperti dalam penelitian Harwell dkk (1996), bahwa ukuran responden dapat mempengaruhi stabilitas estimasi parameter butir.

Berdasarkan panjang tes, akan didapatkan model 3 x 3 x 3 = 27 model data yang dibangkitkan seperti tabel di bawah ini:

Tabel 1

Simulasi 27 model

Responden / Faktor	Panjang Tes								
	10			30			80		
	250	500	1000	250	500	1000	250	500	1000
2F	AA	BA	CA	AA	BA	CA	AA	BA	CA
3F	AB	BB	CB	AB	BB	CB	AB	BB	CB
4F	AC	BC	CC	AC	BC	CC	AC	BC	CC

Keterangan:
Panjang tes terdiri dari 10, 30 dan 80 item
Examinee = 250, 500, dan 1000
Faktor = 2, 3, dan 4
Dari tabel di atas dapat dilihat bahwa peneliti akan melakukan percobaan sebanyak 27 model, yang terdiri dari:

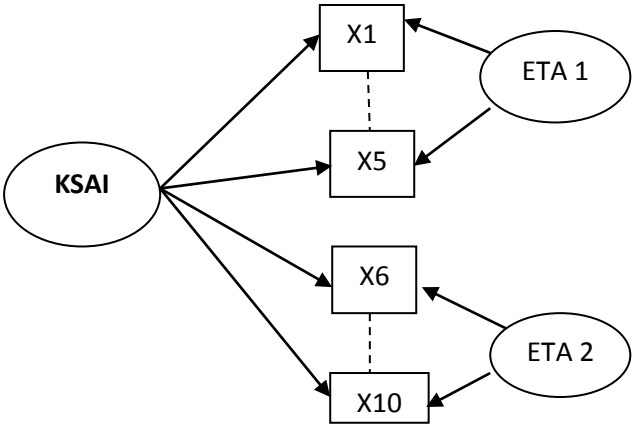
1. Model 10-AA = panjang tes 10, 250 *examinee*, dan 2 faktor
2. Model 10-AB = panjang tes 10, 250 *examinee*, dan 3 faktor
3. Model 10-AC = panjang tes 10, 250 *examinee*, dan 4 faktor
4. Model 10-BA = panjang tes 10, 500 *examinee*, dan 2 faktor
5. Model 10-BB = panjang tes 10, 500 *examinee*, dan 3 faktor
6. Model 10-BC = panjang tes 10, 500 *examinee*, dan 3 faktor
7. Model 10-CA = panjang tes 10, 1000 *examinee*, dan 2 faktor
8. Model 10-CB = panjang tes 10, 1000 *examinee*, dan 3 faktor
9. Model 10-CC = panjang tes 10, 1000 *examinee*, dan 4 faktor
10. Model 30-AA = panjang tes 30, 250 *examinee*, dan 2 faktor
11. Model 30-AB = panjang tes 30, 250 *examinee*, dan 3 faktor
12. Model 30-AC = panjang tes 30, 250 *examinee*, dan 4 faktor
13. Model 30-BA = panjang tes 30, 500 *examinee*, dan 2 faktor
14. Model 30-BB = panjang tes 30, 500 *examinee*, dan 3 faktor
15. Model 30-BC = panjang tes 30, 500 *examinee*, dan 3 faktor
16. Model 30-CA = panjang tes 30, 1000 *examinee*, dan 2 faktor
17. Model 30-CB = panjang tes 30, 1000 *examinee*, dan 3 faktor
18. Model 30-CC = panjang tes 30, 1000 *examinee*, dan 4 faktor
19. Model 80-AA = panjang tes 80, 250 *examinee*, dan 2 faktor
20. Model 80-AB = panjang tes 80, 250 *examinee*, dan 3 faktor
21. Model 80-AC = panjang tes 80, 250 *examinee*, dan 4 faktor
22. Model 80-BA = panjang tes 80, 500 *examinee*, dan 2 faktor
23. Model 80-BB = panjang tes 80, 500 *examinee*, dan 3 faktor
24. Model 80-BC = panjang tes 80, 500 *examinee*, dan 3 faktor
25. Model 80-CA = panjang tes 80, 1000 *examinee*, dan 2 faktor
26. Model 80-CB = panjang tes 80, 1000 *examinee*, dan 3 faktor
27. Model 80-CC = panjang tes 80, 1000 *examinee*, dan 4 faktor

Jumlah Replikasi

Para peneliti telah menggunakan berbagai replikasi, dari 50-21.000 dalam 22 studi direplikasi dalam artikel Mundform (2011). Umumnya, dalam banyak kasus, pemilihan 5000 replikasi sudah cukup menghasilkan hasil yang stabil. Penentuan jumlah replikasi ini mengacu pada prinsipnya yang menyatakan bahwa setiap studi simulasi makin banyak replikasi akan semakin baik, tetapi karena keterbatasan waktu dan terlalu banyaknya kompleksitas dalam menganalisis maka dalam studi ini, peneliti hanya akan melakukan replikasi sebanyak 50 kali. Banyaknya replikasi ini didasarkan pendapat Harwell dkk (1996) bahwa penelitian simulasi untuk terapan teori respons butir, hanya diperlukan sejumlah kecil replikasi, misalnya paling sedikit 10 kali.

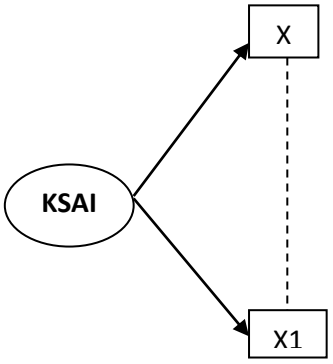
Untuk membangkitkan data atau memperoleh pola respon dari responden yang sesuai dengan parameter yang sebenarnya, diperlukan bantuan komputer. Dalam penelitian ini, program yang digunakan untuk membangkitkan data adalah dengan menggunakan *software* Mplus yang dikembangkan oleh Muthen. Mplus adalah program statistika yang fleksibel untuk menganalisis dengan berbagai pilihan model, estimasi, dan algoritma yang sangat mudah digunakan untuk para penggunanya. Mplus juga memiliki kemampuan yang besar untuk studi MonteCarlo, dimana data yang dibangkitkan langsung dapat dianalisis dengan salah satu model yang diinginkan (Muthen & Muthen, 2010).

Data yang dibangkitkan dalam metode bifaktor di penelitian ini adalah dua, tiga dan empat dimensi atau faktor. Di dalam analisis faktor, dimensi ini dikenal dengan sebutan *etha*. Untuk panjang tes 10, dengan dua dimensi maka ada lima item per dimensi, untuk tiga dimensi maka dibagi menjadi tiga bagian, yaitu lima, tiga dan dua item. Begitu pula dengan panjang tes 10 dengan 4 dimensi, maka dimensinya dibagi menjadi empat bagian, yaitu, 3, 2, 3, 2 item. Untuk lebih jelasnya, berikut peneliti paparkan contohnya dalam gambar berikut:



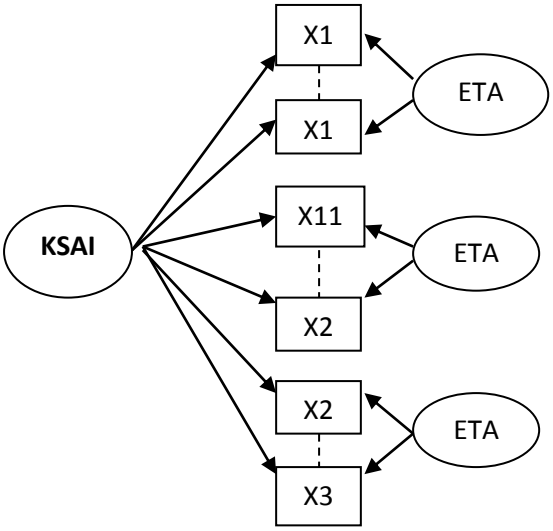
Gambar 1

Gambar Apabila Model dengan 10 Item dan Dibagi Menjadi Dua Dimensi Dilakukan Secara Bifaktor (True)



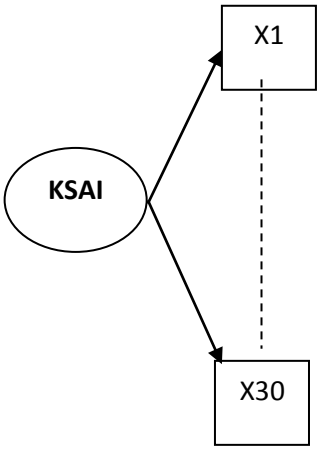
Gambar 2

Gambar Apabila Model dengan 10 Item Dilakukan Secara Unidimensi (Estimate)



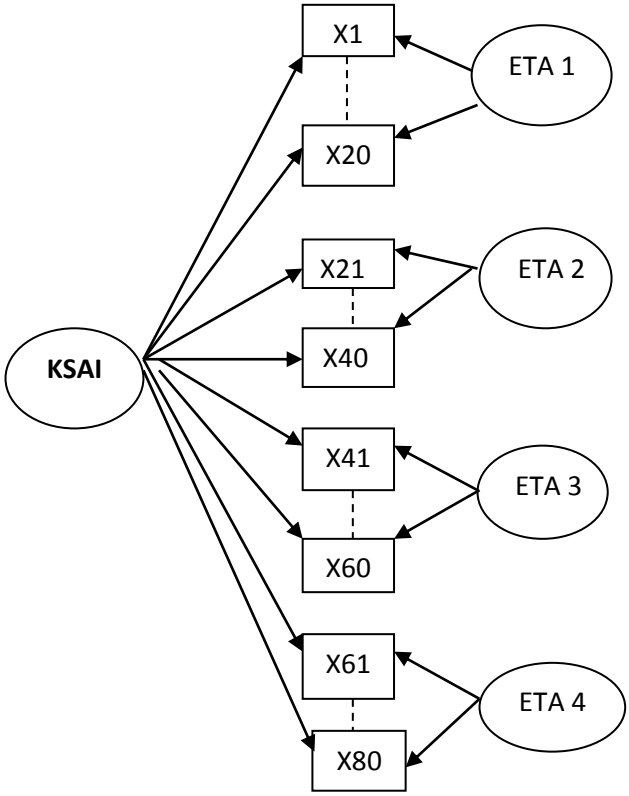
Gambar 3

Gambar Apabila Model dengan 30 Item dan Dibagi Menjadi Tiga Dimensi Diperlakukan Secara Bifaktor (True)



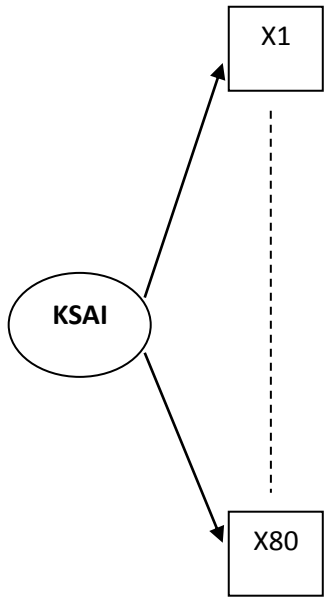
Gambar 4

Gambar Apabila Model dengan 30 Item Diperlakukan Secara Unidimensi (Estimate)



Gambar 5

Gambar Apabila Model dengan 80 Item dan Dibagi Menjadi Empat Dimensi Diperlakukan Secara Bifaktor (True)



Gambar 6

Gambar Apabila Model dengan 80 Item Diperlakukan Secara Unidimensi (Estimate)

Langkah-langkah Penelitian Simulasi

1. Spesifikasi model:
   a. Daya beda = 1, dan kesukaran item = 1,5
   b. Banyaknya faktor 2, 3 dan 4 faktor
   c. Besarnya sampel adalah 250, 500 dan 1000
   d. Banyaknya item adalah 10, 30, dan 80 item
2. Berdasarkan spesifikasi dengan model dibuat matriks korelasi antar item (dengan Mplus)
3. Berdasarkan matriks korelasi tersebut dibangkitkan data dengan Mplus sehingga ada sebanyak 27 jenis data (sesuai dengan kondisi pada tabel 1 sampai 3).
4. Pada setiap kondisi, diestimasi *true score* dua kali:
   a. Diestimasi *true score* pada faktor utama dengan model bifaktor, dan

b. Seluruh item dianggap satu dimensi.

5. Untuk mendapatkan nilai *true score* dari kedua olahan data tersebut, maka peneliti menggunakan *software* Mplus dengan model bias, (adapun sintaknya dapat dilihat di lampiran). Dengan menggunakan model ini, maka pengambilan nilai true pada data tersebut, adalah dengan mengambil nilai median dari masing-masing replikasi. Hasil dari nilai median antara 4b-4a dijadikan kriteria untuk melihat bias (kesalahan estimasi) utama dengan model bifaktor.

HASIL

Setiap model yang dieksperimenkan di dalamnya terdiri dari 50 replikasi dengan jumlah responden yang beragam. Setiap replikasi yang ada akan dianalisis dengan menggunakan *theta true* dan *theta* estmate. Sehingga akan ada dua *theta*. Penghitungan bias atau deviasi pada tiap responden, menggunakan rumus:

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

Keterangan:

$\hat{\theta}$ = *Theta estimate* (hasil analisis menggunakan *as unidimensi*)
θ = *Theta true* (hasil analisis menggunakan data bifaktor)

Dari 27 model yang dilakukan, peneliti mengambil satu contoh bias responden pada replikasi pertama dengan menggunakan model 10-AA, yakni, panjang tes 10, 250 responden, dan 2 faktor, sebagai berikut:

Tabel 2

Nilai Bias Untuk 30 Responden pada Replikasi Pertama Model 10-AA

No. Res	Teta *Estimate*	Teta *True*	Bias	No. Res	Teta *Estimate*	Teta *True*	Bias
res-1	0.613	-0.696	1.309	**res-16**	-0.202	1.297	-1.499
res-2	0.386	0.308	0.078	**res-17**	0.481	-0.852	1.333
res-3	-0.15	1.202	-1.352	**res-18**	0.562	0.501	0.061
res-4	-0.767	-0.517	-0.25	**res-19**	-0.432	-1.863	1.431

No. Res	Teta *Estimate*	Teta *True*	Bias	No. Res	Teta *Estimate*	Teta *True*	Bias
res-5	0.146	-1.755	1.901	**res-20**	-0.185	-1.476	1.291
res-6	-0.617	1.002	-1.619	**res-21**	0.25	-0.127	0.377
res-7	-1.152	1.285	-2.437	**res-22**	-0.526	-0.851	0.325
res-8	0.198	0.326	-0.128	**res-23**	0.216	-1.539	1.755
res-9	-1.463	1.041	-2.504	**res-24**	-0.158	0.583	-0.741
res-10	-0.287	-1.35	1.063	**res-25**	0.512	-0.534	1.046
res-11	-0.069	1.854	-1.923	**res-26**	-0.252	1.007	-1.259
res-12	0.156	1.253	-1.097	**res-27**	-0.178	0.232	-0.41
res-13	-0.167	1.49	-1.657	**res-28**	-0.205	0.378	-0.583
res-14	0.496	0.531	-0.035	**res-29**	0.177	-0.599	0.776
res-15	0.604	0.204	0.4	**res-30**	0.613	-0.696	1.309

Keterangan: model 10-AA, yakni, panjang tes 10, 250 responden, dan 2 faktor.

Sebagaimana dilihat pada tabel 2, setiap responden akan memiliki nilai *theta estimate* dan *theta true*. Gambaran di atas hanya sebagai ilustrasi bagaimana menghitung bias untuk 30 dari 250 responden, sedangkan keadaan sebenarnya perhitungan bias dilakukan kepada semua responden dari responden pertama hingga responden 250.

Dari tabel tersebut juga dapat dilihat bahwa bias yang dihasilkan setiap responden tidak nol artinya terdapat bias atau perbedaan analisis *theta estimate* dan *theta true*. Selanjutnya nilai bias tiap replikasi di atas dapat dihitung *mean*, varian dan standar deviasi untuk setiap replikasi (*within replication*) dan antar replikasi (*over replication*) pada setiap kondisi atau model yang telah ada. Namun bias yang digunakan untuk menghitung *mean*. Varian dan standar deviasi dalam keadaan absolute (tidak diperhitungkan tanda positif dan negatif). Penggunaan bias absolute memiliki kekurangan yakni peneliti tidak mengetahui arah hubungan sedangkan kelebihannya ialah angka yang diperoleh lebih besar, mudah untuk terdeteksi dengan lebih objektif.

Mean, Varian dan Standar Deviasi dari Bias Responden

Bias yang dihasilkan dari selisih nilai *estimate* dengan *true*, dapat diperoleh informasi mengenai *mean*, varian dan standar deviasi untuk setiap replikasi dan antar replikasi dalam sebuah kondisi atau model. Berikut peneliti sajikan hasil

perhitungan *mean*, varian dan standar deviasi untuk satu replikasi dari bias responden

Tabel 3

Mean, Varian, dan Standar Deviasi dari Bias Responden

	10-AA		
Replikasi	***Mean***	**STD**	**Varian**
R-1	-0.03529	1.196357	1.43127

Keterangan:
Model 10-AA, yakni, panjang tes 10, 250 responden, dan 2 faktor.

Dari tabel 3 dapat dilihat nilai *mean* bias pada replikasi pertama model 10-AA sebesar -0.03529, ini berarti secara rata-rata dari 250 responden terdapat bias atau perbedaan antara analisis menggunakan *true score* dan *estimate*. Sedangkan nilai standar deviasi bias responden sebesar 1.196357 yang menunjukkan bias antara responden pertama hingga 250 sangat bervariasi. Hasil analisis tersebut hanya gambaran bagaimana setiap deviasi dari 250 bias responden. Jika dilakukan analisis pada replikasi pertama untuk setiap model yang ada maka hasilnya sebagai berikut:

Tabel 4

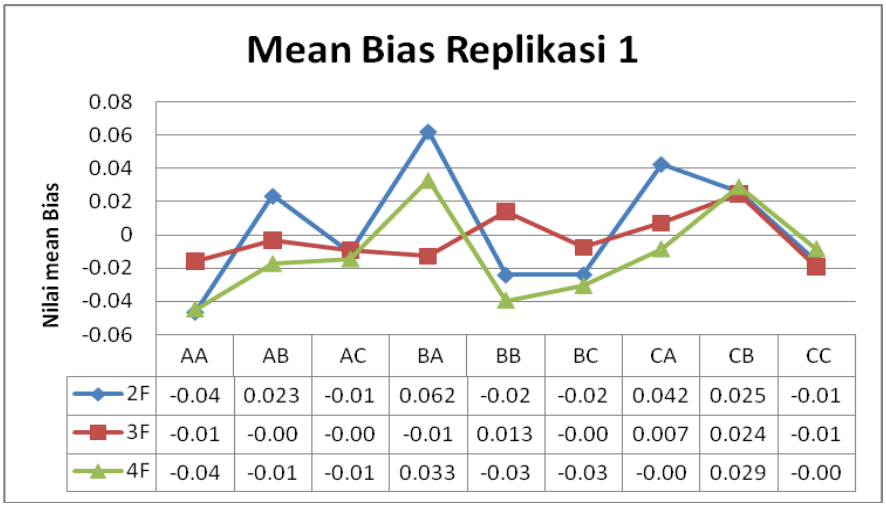
Mean dari Bias Responden Untuk Replikasi Pertama Pada 27 Model

Respon den / Panjang Tes	10			30			80		
Factor	**250**	**500**	**1000**	**250**	**500**	**1000**	**250**	**500**	**1000**
2F	-0.03529	-0.06476	0.034858	-0.05998	0.011784	-0.00566	-0.06916	0.016127	0.018178
3F	0.02354	0.014662	0.030446	-0.03489	0.026902	-0.00216	0.04245	-0.02083	0.015781
4F	-0.01187	0.017475	0.025813	-0.06093	0.052342	-0.00306	0.09643	-0.0429	0.027733

Keterangan:
Panjang tes terdiri dari 10, 30 dan 80 item
Responden = 250, 500, dan 1000 responden
Faktor = 2, 3 dan 4 faktor

Sebelumnya, nilai *mean* yang dihasilkan memang tidak dibulatkan dua angka dibelakang koma, karena perbedaan diantara 27 model yang bervariasi. Dilihat pada table diatas, untuk replikasi pertama pada 27 model terdapat bias antara *theta true* dan *theta estimate*. Hal ini dilihat dari nilai *mean* bias yang dihasilkan lebih besar dari nol.

Namun, untuk keseluruhan model perbedaan ini secara kasat mata tidak terlalu jauh. *Mean* bias responden paling besar menghasilkan nilai 0,052342 terjadi pada model 30 BC, yakni panjang tes 30 item, dengan 4 faktor dan 500 responden. Sedangkan, *mean* bias responden paling kecil dengan nilai -0,09643 terjadi pada model 80 AC yakni dua faktor dengan panjang 80 item, akan lebih jelas bila disajikan dalam bentuk gambar grafik di bawah ini:



	AA	AB	AC	BA	BB	BC	CA	CB	CC
2F	-0.04	0.023	-0.01	0.062	-0.02	-0.02	0.042	0.025	-0.01
3F	-0.01	-0.00	-0.00	-0.01	0.013	-0.00	0.007	0.024	-0.01
4F	-0.04	-0.01	-0.01	0.033	-0.03	-0.03	-0.00	0.029	-0.00

Gambar 7

Mean dari Bias Responden untuk Replikasi Pertama Pada 27 Model

Dari gambar atau grafik 4 dapat dilihat bahwa rata-rata nilai *mean* bias responden antara analisis *true score* dan *true estimate* yang besar terdapat pada item 0,052342. Ini berarti apabila peneliti memiliki data bifaktor, namun dianalisis dengan menggunakan unidimensional, maka hasil *theta* responden akan bias serta tidak memberikan hasil yang sebenarnya mengenai kemampuan responden tersebut.

Kemudian untuk mengetahui apakah bias antara responden satu dengan yang lainnya dalam sebuah replikasi memiliki variasi yang besar atau kecil, peneliti menghitung standar deviasi bias responden pada replikasi pertama dari keseluruhan model. Berikut ini adalah hasil perhitungannya:

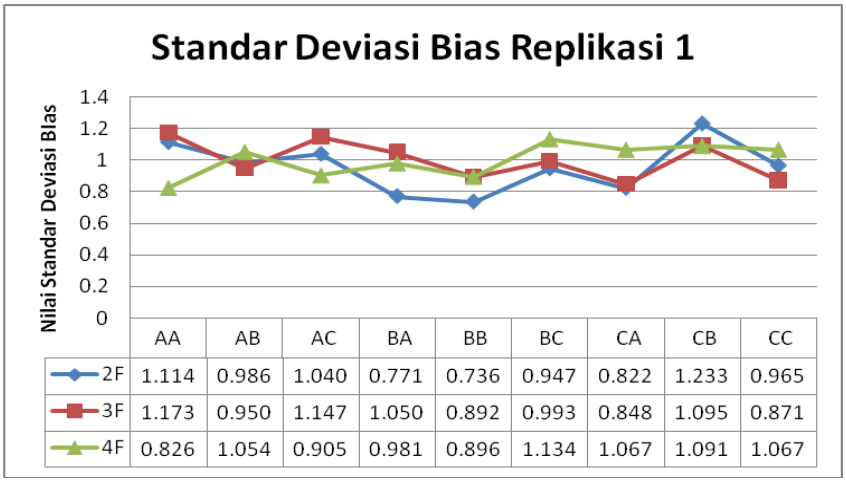
Tabel 5

Standar Deviasi dari Bias Responden untuk Replikasi Pertama pada 27 Model

Respo nden / Fakto r	Panjang Tes								
	10			30			80		
	250	500	1000	250	500	1000	250	500	1000
2F	1,196	0,517	1,149	0,359	0,349	0,335	0,295	0,257	0,247
	3568	3302	1954	62	1184	0678	8217	1926	6261
3F	0,705	0.014	0,674	0,335	0,355	0,336	1,124	0,275	0,250
	3369	662	6688	6403	4299	8897	9949	1355	7647
4F	1,154	0,944	0,431	0,361	0,373	0,333	0,279	0,238	0,224
	541	8368	951	6552	2574	2644	7814	6882	3325

Keterangan:
Panjang tes terdiri dari 10, 30, dan 80 item
Faktor: 2, 3, dan 4

Secara keseluruhan pada tabel diatas, nilai standar deviasi bias responden untuk replikasi pertama antara model satu dengan model yang lainnya tidak terlalu bervariasi. Hal ini bisa dilihat dari nilai standar deviasi berkisar antara 0,014662 hingga 1,1963568. Nilai standar deviasi yang paling besar justru terlihat pada model 10AA (10 item, 250 responden dan 2 faktor), artinya dalam model tersebut bias antara responden pertama hingga responden ke 250 cukup bervariasi sebesar 1,1963568. Sedangkan standar deviasi yang paling kecil terdapat dengan nilai 0,014662 dihasilkan pada model 10 BB (10 item, 500 responden dan 3 faktor). Nilai tersebut menunjukkan bias antar responden tidak terlalu bervariasi. Gambaran mengenai standar deviasi dari model keseluruhan akan lebih jelas bila disajikan dalam bentuk gambar grafik berikut ini:



	AA	AB	AC	BA	BB	BC	CA	CB	CC
2F	1.114	0.986	1.040	0.771	0.736	0.947	0.822	1.233	0.965
3F	1.173	0.950	1.147	1.050	0.892	0.993	0.848	1.095	0.871
4F	0.826	1.054	0.905	0.981	0.896	1.134	1.067	1.091	1.067

Gambar 8

Standar Deviasi Bias Responden untuk Replikasi Pertama pada 27 Model

Interaksi Faktor, Panjang Tes dan Banyaknya Responden

Nilai *mean* bias responden antar replikasi dalam setiap model dapat digunakan untuk mengetahui sejauh apa interaksi dari pengaruh banyaknya faktor, panjang tes dan banyaknya sampel jika dalam nilai yang sama. Berikut hasil penghitungannya:

Tabel 6

Interaksi Faktor, Panjang Tes, dan Banyak Responden

Source	Type III Sum of Squares	df	*Mean* Square	F	Sig.
Corrected Model	1.580[a]	26	0.061	113.403	0
Intercept	0	1	0	0.253	0.615
FACTOR	0.005	2	0.003	4.821	0.008
ITEM	0.409	2	0.204	381.557	.000
RESPONDEN	0.263	2	0.131	245.074	.000
FACTOR * ITEM	0.031	4	0.008	14.623	.000
FACTOR * RESPONDEN	0.025	4	0.006	11.727	.000
ITEM * RESPONDEN	0.8	4	0.2	373.239	.000
FACTOR * ITEM * RESPONDEN	0.047	8	0.006	10.903	.000
Error	0.709	1323	0.001		
Total	2.288	1350			
Corrected Total	2.288	1349			

a. R Squared = .690 (Adjusted R Squared = .684)

Dari tabel di atas dapat diketahui bahwa interaksi antara faktor, panjang tes dan banyaknya responden memiliki nilai R^2 = 0,690 dan nilai signifikansi = 0,000. Artinya, pengaruh faktor, panjang tes dan banyaknya responden bisa meramalkan 69% dari *mean* bias atau perbedaan antara bifaktor dan *as unidimensi* pada data bifaktor. Sedangkan untuk kondisi pengaruh panjang tes dan banyaknya responden dengan jumlah faktor yang disamakan dapat dilihat di bawah ini:

Tabel 7

Interaksi Panjang Tes, Banyaknya Responden, dengan Jumlah Faktor Disamakan (2)

Source	Type III Sum of Squares	Df	*Mean* Square	F	Sig.
Corrected Model	.446[a]	8	0.056	100.609	0
Intercept	0.003	1	0.003	4.557	0.033
ITEM	0.08	2	0.04	71.781	0
SAMPLE	0.095	2	0.047	85.476	0
ITEM * SAMPLE	0.272	4	0.068	122.59	0
Error	0.244	441	0.001		
Total	0.693	450			
Corrected Total	0.691	449			

a. R Squared = .646 (Adjusted R Squared = .640)

Dari tabel di atas dapat diketahui bahwa interaksi antara panjang tes, dan banyaknya responden dengan faktor yang disamakan yakni 2 memiliki R^2 = 0,640 dan nilai signifikansi = 0,000. Dapat dikatakan bahwa pengaruh panjang tes, dan banyaknya responden dengan banyaknya faktor disamakan yakni 2 bisa meramalkan 64% dari *mean* bias atau perbedaan antara unidimensi dengan bifaktor pada data bifaktor.

Tabel 8

Interaksi Panjang Tes, Banyaknya Responden, dengan Jumlah Faktor Disamakan (3)

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	.415a	8	0.052	160.028	0
Intercept	0.002	1	0.002	6.783	0.01
ITEM	0.176	2	0.088	271.539	0
SAMPLE	0.041	2	0.02	63.059	0
ITEM * SAMPLE	0.198	4	0.049	152.757	0
Error	0.143	441	0		
Total	0.56	450			
Corrected Total	0.558	449			

a. R Squared = .744

Dari tabel di atas dapat diketahui bahwa interaksi antara panjang tes, dan banyaknya responden dengan faktor yang disamakan yakni 3 memiliki R^2 = 0,744 dan nilai signifikansi = 0,000. Dapat dikatakan bahwa pengaruh panjang tes, dan banyaknya responden dengan banyaknya faktor disamakan yakni 3 bisa meramalkan 74% dari *mean* bias atau perbedaan antara unidimensi dengan bifaktor pada data bifaktor.

Tabel 9

Interaksi Panjang Tes, Banyaknya Responden, dengan Jumlah Faktor Disamakan (4)

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	.709a	8	0.089	123	0
Intercept	0	1	0	0.502	0.479
ITEM	0.182	2	0.091	126.505	0
SAMPLE	0.153	2	0.077	106.397	0
ITEM * SAMPLE	0.373	4	0.093	129.55	0
Error	0.318	441	0.001		
Total	1.027	450			
Corrected Total	1.026	449			

a. R Squared = .691 (Adjusted R Squared = .685)

Dari tabel di atas dapat diketahui bahwa interaksi antara panjang tes, dan banyaknya responden dengan faktor yang disamakan yakni 4 memiliki R^2 = 0,691 dan nilai signifikansi = 0,000. Dapat dikatakan bahwa pengaruh panjang tes, dan banyaknya responden dengan banyaknya faktor disamakan yakni 4 bisa meramalkan 69% dari *mean* bias atau perbedaan antara unidimensi dengan bifaktor pada data bifaktor.

DISKUSI

Dari penelitian terhadap data bifaktor model, kemudian dianalisis dengan menggunakan *as unidimensi* dan bifaktor diperoleh hasil bahwa terdapat bias

atau perbedaan antara keduanya. Ini dapat diartikan jika suatu keadaan bifaktor, namun hanya dilakukan sekali analisis atau diperlakukan sebagai unidimensi, maka data yang diperoleh sebenarnya tidak menggambarkan keadaan yang sebenarnya.

Dalam simulasi yang sudah dilakukan menggunakan pengaruh faktor, responden dan item, semua replikasi dalam setiap model menghasilkan nilai lebih besar dari nol. Ini artinya setiap replikasi yang dilakukan secara keseluruhan menimbulkan bias antara data yang diperlakukan sebagai unidimensi dan yang diperlakukan sebagai bifaktor (sebagaimana mestinya). Dari ke 27 model percobaan, secara rata-rata nilai *mean* bias paling besar dihasilkan oleh panjang tes 30 item dengan 3 faktor dan 500 responden. Sedangkan nilai *mean* paling kecil terdapat pada panjang tes 80 item, empat faktor dan 250 responden.

Dari penjabaran di atas memiliki arti jika seseorang memiliki data bifaktor tetapi dalam menganalisisnya tetap menganggap sebagai unidimensi atau satu faktor, maka akan menghasilkan bias.

Hasil penghitungan di atas diketahui bahwa terdapat interaksi yang signifikan antara perbedaan faktor, panjang tes dan banyaknya responden yaitu sebesar 69% dari *mean* bias atau perbedaan antara data bifaktor dan data yang diangap sebagai unidimensi. Pengaruh paling besar yang dihasilkan oleh faktor 3 dengan nilai 74%, faktor 4 dengan nilai 69%, dan faktor 2 dengan nilai 64%. Dari sejumlah penjelasan hasil yang telah dirangkumkan diatas, maka didapatkan kesimpulan bahwa hasil yang dilakukan dalam penelitian ini sesuai dengan teori yang disampaikan oleh Humbleton, Swaminathan dan Rogers (2007) bahwa banyaknya faktor, banyaknya responden dan banyaknya item mempengaruhi tingkat korelasinya. Adapun, letak *mean* paling besar dan *mean* paling kecil terletak di nilai tengah, dapat dikarenakan oleh tingkat kesulitan yang cukup besar. Mudah-mudahan artikel ini dapat bermanfaat bagi pembaca.

DAFTAR PUSTAKA

Axelrod, Robert. (2005). *Advancing the art of simulation in the social sciences*. University of Michigan. USA

Chen, Fang Fang, Yiming Jing, Adele Hayes, Jeong Min Lee. (2013) *Two concepts or two approaches: a bifaktor analysis of psychological and subjective well-being.* Department of Psychology. University of Dellaware, Wolf Hall, USA.

Chen, Fang Fang, West, Steephen & Sousa, Karen. (2006). A comparison of bifactor and second order models of quality of life. *Multivariate Behavioral Research*. Lawrence Erlbaum Associates.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologist.* New Jersey: Lawrence Erlbaum Associates. Inc.

Fung, C. (2012). *Ability estimation under different item parameterization and scoring models*. Dissertation, University of North Texas.

Gignac, G. & Watkins, M. (2013). Bifaktor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research, 48:39-662*. Routledge.

Han & Hambleton. (2007). *User's manual for wingen: windows software that generates irt model parameters and item responses*. University of Massachusetts Amherst.

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

Hawell, Michael, Stone, Clement, Tse-Chi Hsu & Levent Kirisci (1996) *Monte carlo studies in* item *response theory*. University of Pittsburgh. *Psychological Measurement vol.20 n0.2, June 1996*.

Jenrich, Robert & Bentler, Peter. (2011). Exploratory bi-faktor analysis. National Institute Of Helath Public Access. *Jurnal Psychometrika,* Oktober 2011.

Kroopnick, Marc Howard. (2010). *Exploring unidimensional proficiency classification accuracy from multidimensional data in a vertical scaling context*. Disertasi dari University of Maryland, College Park.

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

Mundform, D. J., Schaffer, J., Kim, Myoung-Jin, Shaw, D., Thongteeraparp, A., & Supawan, P. (2011). Number of replications required in monte carlo simulation studies: a synthesis of four studies. *Journal of Modern Applied Statistical Methods:* Vol. 10: Iss. 1, Article 4.

Muthen,& Muthen. (2010). *Statistical analysis with latent variables user's guiden.* Los Angeles.

Paxton, Pamela. Et all. (2001). *Monte carlo experiments: design and implementation. Structural Equation Modeling*. Lawrence Erlbaum Associates, Inc.

Reise, Steven. (2012). The rediscovery of bifaktor measurement models. *Journal Multivariate Behavioral Research.* Oct. 2012. Vol 47.

Raychaudhuri, Samik (2008). *Introduction to monte carlo simulation*. Proceedings of the 2008 Winter Simulation Conference. USA.

Retnawati, Heri. (2008*). Estimasi efisiensi relatif tes berdasarkan teori respons butir dan teori tes klasik*. Disertasi. Yogyakarta: Program Pascasarjana Universitas Negeri Yogyakarta.

Simon, Mayuko Kanada. (2008). *Comparison of concurrent and separate multidimensional IRT lingking of item parameters.* University of Minnesota. Tesis.

Sinharay, dan Haberman. (2010). Reporting of subscores using muldimensional item response theory. *Journal Psikometrika* col. 75.no. 2 hal. 209-227.

Seranno, Daniel. (2010). *A second order growth model for longitudinal item response data.* Disertasi University of Nourth Carolina.

Umar, Jahja. (2012a). *Penilaian dan peningkatan mutu pendidikan di indonesia*. Ciputat: UIN Press.

Umar, Jahja. (2012b). Mengenal lebih dekat konsep reliabilitas skor tes. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia.* April. 2012. Vol.II.

Widhiarso, Wahyu. (2011). *Teori pengukuran*. Bahan Materi Fakultas Psikologi UGM.