INVARIANSI SEBAGAI BUKTI VALIDITAS PENGUKURAN

Ali Ridho
UIN Maliki Malang
ali.ridho@live.com

Abstrak. Tantangan yang dihadapi sebuah pengukuran adalah bagaimana tes menghasilkan skor yang presisi sepanjang skala pengukuran, komparabel, dan adil. Konsepsi validitas dalam pengukuran pendidikan dan psikologi mulai jelas sejak muncul tulisan Cronbach dan Meehl (1955) yang mengupas 4 jenis validitas dalam pengukuran. Keempat validitas itu adalah (1) validitas prediktif (*predictive validity*), (2) validitas konkuren (*concurrent validity*), (3) validitas isi (*content validity*), dan (4) validitas konstrak (*construct validity*). Selain membahas beberapa irelevansi pengategorian semacam ini, dalam tulisan ini dibahas pula pengertian mengenai validitas dalam konteks terkini, khususnya tentang invariansi (ekuivalensi) konstrak antar kelompok pada tes dengan skala besar (*large scale testing*). Diskusi dalam pembahasan disertai dengan ilustrasi menggunakan data respons peserta Seleksi Penerimaan Mahasiswa Baru (SPMB) dalam uji struktur dimensi (faktor) yang dimodelkan melalui *Multigroup Nonlinear Confirmatory Factor Analysis* (MGNCFA).

***Kata kunci**: validitas, ekuivalensi/invariansi, Multigroup Nonlinear Confirmatory Factor Analysis (MGNCFA)*

A. Pendahuluan

Pengukuran merupakan kegiatan yang lazim, bahkan rutin dilakukan dalam dunia pendidikan dan psikologi. Berdasarkan pengamatan penulis, banyak laporan penelitian di bidang pendidikan, psikologi, dan ekonomi yang menggunakan validitas aitem sebagai justifikasi valid tidaknya skor yang dihasilkan oleh alat ukur yang digunakan dalam penelitian dalam bentuk korelasi aitem-total (r_{it}) atau korelasi aitem-total terkoreksi (*corrected item-total correlation*, $r_{i(t-i)}$). Hal tersebut perlu dikoreksi karena r_{it} bukan menunjukkan bukti valid atau tidaknya skor yang dihasilkan oleh alat ukur. Kenyataan ini pernah disinggung oleh Naga (2004) yang mengatakan bahwa persoalan validitas tidak sesederhana itu, namun proses validasinya dilakukan terhadap hasil ukur sehingga bisa membuktikan konstrak yang dikembangkan betul-betul berlaku pada subjek yang menjadi tujuan ukur.

Dalam praktik di dunia pendidikan dan psikologi, sering kali peneliti melakukan pembandingan antar kelompok dalam atribut laten tertentu. Hal yang penting sebelum membandingkan atribut laten pada kelompok yang berbeda adalah menguji berlakunya konstrak, apakah berlaku secara ekuivalen pada kelompok-kelompok subjek ukur yang terlibat. Pengujian ini merupakan prasyarat bilamana kelompok yang berbeda akan dibandingkan (Chen, 2007). Kelompok ini dapat diidentifikasi berdasarkan kultur, etnis, gender, umur, dan kontrol versus eksperimen. Bila ekuivalensi belum ditegakkan sementara pembandingan dilakukan, ini sama dengan membandingkan dua pengukuran dalam skala belum setara. Lebih jauh, bila pengujian perbedaan rata-rata skor antar kelompok tidak didahului oleh ekuivalensi konstrak, kesimpulan yang dihasilkan tidak akan akurat (Vandenberg, 2002), perbedaan yang teramati bisa jadi menunjukkan ketidaksetaraan pengukuran dan bukan perbedaan yang sebenarnya dalam konstrak yang dibandingkan (Ployhart & Oswald, 2004). Oleh sebab itu penting kiranya menegakkan ekuivalensi berlakunya konstrak yang diukur pada kelompok-kelompok yang hendak dibandingkan.

Sejalan dengan pendapat Gorin (2007), Embretson (2007), Sireci (2007), penting kiranya menemukan bukti-bukti secara internal bahwa sebuah tes betul-betul valid mengukur selaras dengan fondasi konstraknya. Dengan kata lain, struktur internal tes yang berupa aitem-aitem perlu ditemukan tingkat validitasnya. Aspek struktural dalam validitas dapat ditegakkan

dengan cara *Confirmatory Factor Analysis* (CFA) (Dimitrov, 2010). Tulisan ini mengupas tentang *measurement equivalent/invariance* (ME/I), khususnya *Multigroup Nonlinear Confirmatory Factor Analysis* (MGNCFA). Ilustrasi penerapan dipaparkan menggunakan data respons peserta Seleksi Penerimaan Mahasiswa Baru (SPMB).

B. Kajian Pustaka

Hakikat Korelasi Aitem-Total

Korelasi aitem-total (r_{it}) dianggap oleh sebagian besar praktisi pendidikan dan psikologi sebagai validitas karena selama ini pemilihan aitem-aitem pada saat mengembangkan alat tes lebih banyak didasarkan pada koefisien ini. Melalui koefisien ini pula sebuah aitem sebuah aitem dipertahankan menjadi bagian dari suatu alat tes, namun bisa juga dibuang, diperbaiki atau diganti. Oleh karena pentingnya peranan r_{it} dalam praktik pengembangan aitem itulah maka koefisien ini sering disebut sebagai validitas aitem.

Penggunaan istilah validitas aitem pada r_{it} sebenarnya tidak tepat. Makna yang terkandung dalam koefisien ini yaitu bahwa bila sebuah aitem memiliki r_{it} tinggi maka skor aitem memiliki nilai yang sejalan dengan nilai total; demikian pula sebaliknya bila sebuah butir memiliki r_{it} rendah maka skor aitem tidak sejalan dengan nilai total. Makin tinggi r_{it} berarti sebuah aitem makin mampu mengelompokkan mana saja peserta tes yang memiliki atribut ukur yang tinggi dan peserta tes mana saja yang memiliki atribut ukur rendah. Oleh sebab itu koefisien r_{it} pada hakikatnya memberikan informasi tentang *daya beda* aitem, bukan validitas aitem.

Kalimat lain yang dapat digunakan untuk mendeskripsikan makna r_{it} adalah; konsistensi skor aitem dengan skor total. Oleh karena itu tidak mengherankan manakala peneliti berkeinginan memperoleh reliabilitas konsistensi internal α Cronbach, mereka menjatuhkan pilihan pada aitem-aitem dengan r_{it} tinggi. Makin banyak aitem-aitem dalam tes yang memiliki r_{it} tinggi maka makin tinggi pula reliabilitas konsistensi internal α Cronbach yang diperoleh. Dengan demikian dapat dikatakan bahwa besarnya rata-rata r_{it} berbanding lurus dengan konsistensi internal α Cronbach. Bukti-bukti kajian psikometrik kenyataan ini dapat ditelusuri dalam tulisan Naga (2004).

Konsepsi Validitas

Konsepsi validitas dalam pengukuran pendidikan dan psikologi mulai jelas sejak muncul tulisan Cronbach dan Meehl (1955) yang mengupas 4 jenis validitas dalam pengukuran. Keempat validitas itu adalah (1) validitas prediktif (*predictive validity*), (2) validitas konkuren (*concurrent validity*), (3) validitas isi (*content validity*), dan (4) validitas kontrak (*construct validity*). Baik validitas prediktif ataupun konkuren, keduanya dapat disebut sebagai validitas kriteria. Validitas isi ditegakkan dengan terwakilinya domain yang hendak diukur melalui sampel aitem yang representatif. Validitas konstrak menunjukkan sejauh mana sebuah tes telah mengukur apa yang memang hendak diukur. Untuk itu dibutuhkan bukti-bukti yang memadai. Hal ini tercermin dalam poin ke 4 (empat) pada kesimpulan mereka diakhir artikel tersebut:

> *Many types of evidence are relevant to construct validity, including content validity, interitem correlations, intertest correlations, test-"criterion" correlations, studies of stability over time, and stability under experimental intervention. High correlations and high stability may constitute either favorable or unfavorable evidence for the proposed interpretation, depending on the theory surrounding the construct.*

Empat puluh (40) tahun semenjak pendapat Cronbach dan Meehl (1955) dipublikasikan, barulah muncul pendapat Messick (1995) yang mendefinisikan validitas sebagai satu kesatuan, tidak terbagi-bagi sebagaimana menurut

Cronbach dan Meehl. Messick mengemukakan bahwa validitas adalah ringkasan evaluatif baik dalam bentuk bukti atau konsekuensi pada interpretasi dan penggunaan skor hasil tes. Konsep pengertian validitas yang dikemukakan Messick ini kemudian menjadi referensi yang diacu dalam standar bersama antara *American Psychological Association* (APA), *American Educational Research Association* (AERA) dan *National Council on Measurement in Education* (NCME) dalam mendefinisikan validitas; ialah: *Validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of test* (APA, AERA, & NCME, 1999). Sejalan dengan pengertian ini, *Educational testing Service* (ETS) memberikan pengertian: "*Validity is the extent to which inferences and actions made on the basis of a set of scores are appropriate and justified by evidence. Validity refers to how the scores are used rather than to the assessment itself*" (ETS, 2002).

Kata kunci dalam memahami pengertian validitas adalah pada sejauh mana ketepatan interpretasi skor yang dihasilkan oleh tes bersesuaian dengan tujuan ukurnya. Interpretasi dan penggunaan skor haruslah mengacu pada konstrak yang memang sejak awal sengaja dikembangkan untuk diukur oleh tes sebagai alatnya. Bila tujuan alat tes adalah mengungkap atribut psikologis tertentu – misalnya kejujuran, *self-esteem*, kepuasan menikah, intelegensi– maka bagaimana interpretasi terhadap skor yang dihasilkan oleh tes dikaitkan dengan atribut yang hendak diukur tersebut. Selain itu, interpretasi ini harus dilandaskan pada bukti-bukti ilmiah secara teoritik (*content related evidence*) dan empirik (*empirical related evidence*).

Interpretasi skor seharusnya tidak hanya mengacu pada penekanan yang dimaksudkan oleh tes, namun pada norma sosial pula. Ini adalah tuntutan yang logis sebab kondisi fisik dan psikologis sekelompok subjek tertentu akan terkait selalu terkait dengan lingkungan serta norma sosial yang berlaku di sana. Lingkungan sosial dan norma yang disinyalir dapat menjadikan sebuah konstrak –yang dioperasionalkan dalam bentuk tes– direspon dengan cara yang berbeda oleh kelompok-kelompok yang berbeda. Kelompok-kelompok ini dapat dikategorikan berdasarkan berbagai kriteria, misalnya kelompok perempuan-lelaki, kota-desa, bahkan lintas negara atau benua. Oleh sebab itu kesesuaian antara sebuah konstrak hipotetis yang dituangkan dalam sebuah alat ukur hendaknya dibuktikan pula dengan bukti-bukti bahwa konstrak tersebut berlaku secara umum (*general*) pada sampel yang relevan dengan tujuan alat ukur. Sebuah alat ukur bisa jadi mengungkap atribut yang memiliki struktur tertentu bila dikenakan pada kelompok sampel tertentu namun memiliki struktur yang berbeda pada sampel lain. Disinilah pentingnya kesetaraan invariansi struktur konstrak pada kelompok sampel dalam populasi yang menjadi subjek ukur tes; dikenal dengan istilah Measurement Equivalence / Invariance (ME/I).

Measurement Equivalence / Invariance (ME/I)

Hal yang penting sebelum membandingkan atribut psikologis pada kelompok yang berbeda adalah menguji berlakunya konstrak, apakah berlaku secara ekuivalen pada kelompok-kelompok subjek ukur yang terlibat. Pengujian ini merupakan prasyarat bilamana kelompok yang berbeda akan dibandingkan (Chen, 2007). Kelompok ini dapat diidentifikasi berdasarkan kultur, etnis, gender, umur, dan kontrol versus eksperimen. Bila ekuivalensi belum ditegakkan sementara pembandingan dilakukan, ini sama dengan membandingkan dua pengukuran dalam skala yang berbeda. Bila pengujian perbedaan rerata antar kelompok tidak didahului oleh ekuivalensi konstrak, kesimpulan yang dihasilkan tidak akan akurat (Vandenberg, 2002), perbedaan yang teramati bisa jadi menunjukkan ketidaksetaraan pengukuran dan bukan perbedaan yang sebenarnya dalam konstrak yang dibandingkan (Ployhart & Oswald,

2004). Oleh sebab itu penting kiranya menegakkan ekuivalensi berlakunya konstrak yang diukur pada kelompok-kelompok yang hendak dibandingkan.

Ekuivalensi konstrak merupakan aspek yang penting ditegakkan dalam tes sebagai instrumen pengukuran. Bila tes ditujukan untuk diadministrasikan pada populasi yang bersifat heterogen, harus ditegakkan terlebih dulu karakteristik psikometrik yang ekuivalen antar kelompok dalam populasi. Sebuah aitem dapat bersifat bias manakala aitem tidak dapat berfungsi sama antar kelompok. Lebih serius lagi pada level tes. Dalam situasi dimana kemampuan yang diungkap tes akan menentukan masa depan subjek (*high stakes testing*), persoalan invariansi pengukuran adalah isu yang penting untuk ditegakkan. Guna menjamin keberlakuan konstrak yang diungkap oleh sebuah tes pada kelompok-kelompok dalam populasi subjek, perlu diselidiki keberlakuannya.

Dalam bidang pengukuran, istilah bagi berlakunya konstrak pada kelompok-kelompok yang berbeda disebut sebagai invariansi atau ekuivalensi pengukuran (*measurement equivalence or invariance*, ME/I) (Jones-Farmer, 2010). ME/I mengacu pada pengukuran konstrak atau sehimpunan konstrak yang setara dalam dua atau lebih kelompok atau antar waktu melalui satu set variabel atau indikator amatan (Bauer, 2005). Variabel ataupun indikator amatan ini dikenal pula dengan istilah variabel manifes.

Pengujian ME/I sebenarnya sudah banyak dilakukan para peneliti menggunakan prosedur *Multigroup* CFA (MGCFA). Terdapat 49 artikel yang teridentifikasi terbit tahun 2009 s.d. 2011; enam diantaranya ditulis oleh Booth, Irwing, dan Booth (2011), Reniers, Corcoran, Drake, Shryane, dan Völlm (2011), Bowden, Saklofske, dan Weiss (2011), Nair, White, Knight, dan Roosa (2009), South, Krueger, dan Iacono (2009), serta Rotgans dan Schmidt (2009). Seluruh artikel tersebut menggunakan MGCFA dalam model yang bersifat linier.

Nonlinear Confirmatory Factor Analysis (NCFA)

Sebelum membahas tentang *nonlinear confirmatory factor analysis* (NCFA), berikut ini akan dikemukakan sekilas tentang *linear confirmatory factor analysis* (LCFA). Hal ini perlu penulis kemukakan karena sebagian besar peneliti bidang pendidikan dan psikologi lebih familier dengan NCFA. Dari sisi tujuan dan konsepsi, banyak persamaan antara LCFA dan NCFA (Flowers, Raju, & Oshima, 2002; Maydeu-Olivares, Hernández, & McDonald, 2006; Reise, Widaman, & Pugh, 1993). Perbedaannya adalah LCFA didasarkan pada fungsi linier, sementara NCFA nonlinier. Oleh sebab itu dalam tulisan ini akan dikemukakan istilah dan konsepsi LCFA sebagai pengantar ke model NCFA.

Tujuan LCFA adalah mengidentifikasi faktor laten yang dapat menjelaskan variasi dan kovariasi antar indikator (Brown, 2006). Sebuah indikator ini dalam implementasinya dapat berupa satu atau sekumpulan aitem. LCFA dimulai dengan mendefinisikan variabel laten yang akan diukur berdasarkan teori atau pengetahuan terdahulu (Jöreskog, 2007). Dalam LCFA peneliti harus menentukan terlebih dahulu semua aspek dalam model faktor: jumlah faktor, pola muatan faktor, dan seterusnya.

Apabila melalui prosedur LCFA kemudian terbukti secara meyakinkan bahwa struktur konstrak yang dihasilkan oleh alat ukur membentuk hierarki sebagaimana argumentasi teoritik yang mendasarinya, hal tersebut adalah bukti awal bahwa konstrak yang dikembangkan telah berlaku dalam populasi subjek yang menjadi tujuan. Bukti validitas perlu ditingkatkan lagi manakala akan dilakukan pembandingan terhadap kelompok-kelompok yang merupakan bagian dari populasi tersebut melalui penegakan keberlakuan konstrak yang sama antar kelompok tersebut.

Keberlakuan konstrak yang sama antar kelompok disebut dengan ekuivalensi konstrak. Ekuivalensi konstrak merupakan aspek yang penting dalam pengembangan tes. Bila tes ditujukan untuk diadministrasikan pada populasi yang

bersifat heterogen, harus ditegakkan terlebih dulu karakteristik psikometrik yang ekuivalen antar kelompok dalam populasi. Sebuah aitem dapat bersifat bias manakala aitem tidak dapat berfungsi sama antar kelompok. Lebih serius lagi pada level tes, dalam situasi dimana kemampuan yang diungkap tes akan menentukan masa depan subjek (*high stakes testing*), persoalan bias tes adalah isu yang penting untuk ditegakkan. Guna menjamin keberlakuan konstrak yang diungkap oleh sebuah tes pada kelompok-kelompok dalam populasi subjek, perlu diselidiki keberlakuannya.

Dalam bidang pengukuran, istilah bagi berlakunya konstrak pada kelompok-kelompok yang berbeda disebut sebagai invariansi atau ekuivalensi pengukuran (*measurement equivalence or invariance*, ME/I) (Jones-Farmer, 2010). ME/I mengacu pada pengukuran konstrak atau sehimpunan konstrak yang setara dalam dua atau lebih kelompok atau antar waktu melalui satu set variabel atau indikator amatan (Bauer, 2005). Variabel ataupun indikator amatan ini dikenal pula dengan istilah variabel manifes yang, dalam konteks pengukuran, dapat berupa aitem-aitem tes.

Model LCFA terdiri dari muatan faktor, varians unik, dan varians faktor. Muatan faktor merupakan kemiringan regresi (prediksi) indikator (aitem) dari faktor laten. Varians unik adalah varians indikator yang tidak dapat dijelaskan oleh faktor laten. Varians unik dikenal pula sebagai "varians eror" dan "ketidakreliabelan".

Persamaan dasar model faktor bersama dalam LCFA adalah:

$$x_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \ldots + \lambda_{im}\xi_m + \varepsilon_i \qquad (1)$$

dimana x_i adalah indikator ke-i, λ_{im} menunjukkan muatan faktor indikator i pada faktor ξ ke-m. Persamaan ini dapat diringkas dalam sebuah persamaan yang mengekspresikan hubungan antara variabel amatan (x), faktor laten (ξ), dan varians unik (ε):

$$x = \Lambda_x \xi + \varepsilon \qquad (2)$$

atau dalam bentuk matriks:

$$\mathbf{\Sigma} = \mathbf{\Lambda}_x \mathbf{\Phi} \mathbf{\Lambda}_x^{'} + \mathbf{\Theta}_\delta \qquad (3)$$

dimana $\mathbf{\Sigma}$ adalah matriks korelasi $p \times p$ dari p aitem, $\mathbf{\Lambda_x}$ adalah matriks $p \times m$ dari muatan faktor λ, $\mathbf{\Phi}$ adalah matriks simetris $m \times m$ dan $\mathbf{\Theta_\varepsilon}$ adalah matriks diagonal varians unik $p \times p$.

Beralih pada NCFA yang bersesuaian dengan *item response theory* (IRT), parameter yang dihasilkan oleh persamaan (1), (2) dan (3) dapat ditransformasikan menjadi parameter-parameter aitem (McDonald, 1999). Daya beda (a) dan kesukaran (b) dapat diperoleh dengan:

$$a_i = \frac{\lambda_i}{\sqrt{1 - \lambda_i' \varphi \lambda_i}} \qquad (4)$$

$$b_i = \frac{-\tau_i}{\sqrt{1 - \lambda_i' \varphi \lambda_i}} \qquad (5)$$

dimana λ_i adalah vektor muatan faktor aitem i, φ adalah matriks kovarians faktor, dan τ_i adalah nilai batas (*treshold*) aitem i.

Bila ditambahkan faktor penskalaan sebesar 1,7 pada persamaan **Error! Reference source not found.** dan **Error! Reference source not found.** bentuk permukaan yang dihasilkan oleh dua fungsi tersebut akan mendekati model *multidimensional normal ogive* (MNO). Dengan melandaskan pada model MNO, McDonald (2000) mengemukakan bahwa untuk sehimpunan p respons biner (dikotomi) U_1, U_2, …, U_p, yang berharga 0 dan 1, probabilitas menjawab benar dapat didefinisikan sebagai:

$$P(U_i = 1 \mid \theta) = N(\beta_{i0} + \beta_i' \theta) = N(\beta_{i0} + \beta_{i1}\theta_1 + \beta_{i2}\theta_2 + \ldots + \beta_{im}\theta_m) \qquad (6)$$

dimana $N(\cdot)$ merupakan fungsi distribusi kumulatif normal, m adalah dimensi ruang laten (banyaknya θ atau atribut laten).

Christoffersson (dalam McDonald, 1999) membuat formulasi yang ekuivalen dengan persamaan (6), yaitu

$$V_i = \boldsymbol{\lambda}_i' \theta + \varepsilon_i, \qquad (7)$$

dimana $\mathbf{\Lambda}' = [\lambda_1, \lambda_2, ..., \lambda_n]$ adalah matriks muatan faktor bersama, **θ** merupakan vektor atribut laten, ε_i adalah varians unik faktor ke-*i*, dengan asumsi:

$$U_i = \begin{cases} 1 & \text{jika } V_i \geq \tau_i \\ 0 & \text{jika } V_i < \tau_i \end{cases}. \quad (8)$$

dimana τ_i adalah parameter batas aitem.

Proporsi menjawab benar sebuah aitem *i*, ditunjukkan dengan

$$\pi_i = N(-\tau_i), \quad (9)$$

dan proporsi bersama para peserta yang menjawab benar aitem *i* dan *k* dapat diekspresikan dalam fungsi tetrakorik. Hasilnya ialah parameterisasi model MNO menjadi "analisis faktor aitem" :

$$P(U_i = 1 | \theta) = N\{(\lambda_i'\theta - \tau_i)\psi_{ii}^{-1/2}\}. \quad (10)$$

Simbol λ_i adalah muatan faktor terstandar, ψ_{ii} adalah varians residu (varians unik).

Model dalam persamaan (7) dapat diformulasikan menjadi

$$\mathbf{v} = \mathbf{\Lambda\theta} + \boldsymbol{\varepsilon} \quad (11)$$

dimana **v** adalah variabel kontinu V_1, V_1, …, V_n. Berdasarkan formula pada persamaan (11), ditentukan banyaknya dimensi (θ) terlebih dahulu sehingga model dapat diidentifikasi dan akhirnya dapat dimaknai.

Metode *full information maximum likelihood* akan menghasilkan χ^2 yang dapat digunakan untuk menghitung indeks *Goodness-of-fit* (GFI), *d*, dan *root mean Square error approximation* (RMSEA) (McDonald, 2000)

$$d = (\chi^2 - \mathrm{DF}) / n \quad (12)$$

$$\mathrm{RMSEA} = \sqrt{(d / \mathrm{DF})} \quad (13)$$

Selain itu, dapat pula digunakan ukuran kecocokan data γ_{ULS}

$$\gamma_{ULS} = 1 - \frac{\mathrm{Tr}(\mathbf{R}^2)}{\mathrm{Tr}(\mathbf{S}^2)}, \quad (14)$$

dimana **S** adalah matriks korelasi proporsi bersama antar pasang aitem, **R** adalah matriks residu.

Multigroup Nonlinear Confirmatory Factor Analysis (MGNCFA)

Sejalan dengan pendapat Gorin (2007), Embretson (2007), Sireci (2007), penting kiranya menemukan bukti-bukti secara internal bahwa sebuah tes betul-betul valid mengukur selaras dengan fondasi konstraknya. Dengan kata lain, struktur internal tes yang berupa aitem-aitem perlu ditemukan tingkat validitasnya. Aspek kesetaraan struktural dalam validitas dapat ditegakkan dengan cara *Multigroup Confirmatory Factor Analysis* (MGCFA) (Dimitrov, 2010).

Dalam bidang pengukuran, istilah bagi berlakunya konstrak pada kelompok-kelompok yang berbeda disebut sebagai invariansi atau ekuivalensi pengukuran (*measurement equivalence or invariance*, ME/I) (Jones-Farmer, 2010). ME/I mengacu pada pengukuran konstrak atau sehimpunan konstrak yang setara dalam dua atau lebih kelompok atau antar waktu melalui satu set variabel atau indikator amatan (Bauer, 2005). Prosedur MGCFA linier ini sudah banyak digunakan untuk meneliti *measurement equivalence/invariance* (ME/I) (misalnya Beaujean, McGlaughlin, & Margulies, 2009; Booth dkk., 2011; Immekus & Maller, 2010; Molenaar, 2009).

Persyaratan utama dalam invariansi pengukuran adalah bahwa skor manifes harapan seseorang yang memiliki kemampuan laten tertentu, bersifat independen dari keanggotaan kelompok. Katakanlah seorang laki-laki dan perempuan memiliki kemampuan yang sama dalam hal matematika. Perbedaan sistematik skor amatan mereka dalam sebuah tes matematika mengisyaratkan adanya bias terkait dengan gender. Katakanlah $P(\boldsymbol{\theta})$, yaitu probabilitas

menjawab benar pada aitem, dan vektor **θ** menunjukkan variabel laten yang mendasari skor $P(\boldsymbol{\theta})$. $P(\boldsymbol{\theta})$ seharusnya hanya tergantung dari kemampuan laten, tanpa ada sumbangan kelompok, misalnya gender.

Dengan memasukkan variabel kelompok ke dalam persamaan (6) maka jika invariansi dapat ditegakkan, dapat ditentukan lokasi vektor **θ** dan skor harapan seharusnya bernilai sama antara laki-laki (lk) dan perempuan (pr), akan diperoleh formulasi

$$P(U_i = 1 \mid \boldsymbol{\theta}, \text{lk}) = P(U_i = 1 \mid \boldsymbol{\theta}, \text{pr}) = P(U_i = 1 \mid \boldsymbol{\theta}) \quad (15)$$

Pada MGCFA linier, ME/I dapat dikategorikan menjadi 3 (tiga): konfigural, pengukuran (muatan faktor), dan struktural (Brown, 2006; Dimitrov, 2010), demikian juga dalam konteks nonlinier IRT (Reise dkk., 1993). Bila yang diuji ME/I dari sudut pandang MGCFA adalah muatan faktor dan *intercept*, dalam sudut pandang IRT, parameter yang diperhitungkan adalah daya beda (*a*) dan tingkat kesukaran (*b*).

Berdasarkan banyaknya dimensi yang telah teridentifikasi melalui prosedur uji dimensionalitas, justifikasi ahli, dan atau pengembang tes, dapat ditentukan vektor **θ** yang mendasari peserta tes menjawab benar pada sebuah aitem. Struktur model yang menyusun pun dapat diuji secara sendiri pada populasi tertentu, ataupun terpisah pada beberapa kelompok. Persoalan skala antar kelompok, hal tersebut tidaklah menjadi masalah karena estimasi parameter aitem dapat dilakukan secara simultan beberapa kelompok sekaligus (Yao & Li, 2010).

C. Metode

Data penelitian ini skor 28 aitem yang telah terpilih memiliki $r_{it} \geq 0{,}25$ pada Subtes Verbal Potensi Akademik (PA) Seleksi Penerimaan Mahasiswa Baru tahun 2012. Aitem-aitem tersebut direspons oleh 54000 peserta. Peneliti mengacak secara random sehingga diperoleh sampel dengan ukuran 14000 (7000 perempuan dan 7000 lelaki).

Untuk menguji secara konfirmatori pada struktur model aitem-aitem dalam *multidimensional item response theory* (MIRT), dapat digunakan metode *adaptive quadrature* (diaplikasikan dalam MPlus dan IRTPRO), *bayesian markov chain monte carlo*, *maximum likelihood estimation*, ataupun *limited-information least square* (diaplikasikan dalam NOHARM) (Cai, Yang, & Hansen, 2011). Selain itu, dalam menguji ME/I antar kelompok dalam kerangka MIRT, bisa juga digunakan metode *bayesian markov chain monte carlo* yang telah diaplikasikan dalam software *Bayesian Multivariate Item Response Theory*, disingkat BMIRT (Yao, 2010). Pada penelitian ini, penulis menggunakan metode *adaptive quadrature*.

D. Diskusi dan Pembahasan

Guna mengetahui apakah struktur model yang diterapkan dapat mengakomodir data respons maka diperlukan evaluasi kecocokan model. Menurut Whittaker, Chang, dan Dodd (2012), sampai dengan saat ini ada beberapa kriteria kecocokan data yang dapat digunakan dalam menentukan model mana yang paling dapat menjelaskan data, yaitu: *likelihood ratio test* (LRT) atau perbedaan -2loglikelihood, *Akaike's Information Criterion* (AIC), *Bayesian Information Criterion* (BIC), *Cross-validation Log Like*lihood (CVLL), dan *Deviance Information Criterion* (DIC). Mereka menyatakan bahwa BIC efektif untuk memilih indeks kecocokan data untuk sampel yang besar ($n > 1000$).

Tabel 1. Perbandingan Kecocokan Model MIRT M1PL dan M2PL

Model	-2loglikelihood	AIC	BIC
M1PL	419891,80	419951,80	420178,20
M2PL	416388,40	416556,40	417190,33
D	3503,40	3395,40	2987,87

Keterangan: M1PL = model logistik 1 parameter multidimensi, M2PL = model logistik 2 parameter multidimensi, D = nilai perbedaan kecocokan data, AIC = *Akaike's*

Information Criterion, BIC = *Bayesian Information Criterion*.

Model awal yang diterapkan untuk mengestimasi parameter-parameter yang melekat pada aitem-aitem dalam alat ukur yaitu model logistik multidimensi 2 (dua) parameter (M2PL) dan model logistik multidimensi 1 (satu) parameter (M1PL) dalam kerangka IRT multidimensi. Pada analisis ini, data keseluruhan (14000) dianalisis secara bersama-sama tanpa memperhatikan kelompok perempuan ataupun lelaki. Hasil analisis kecocokan data dengan model-model MIRT antara M1PL dan M2PL disajikan pada **Tabel 1**. Pada tabel tersebut dapat dilihat perbedaan (D) -2loglikelihood, AIC, dan BIC. Ketiga perbedaan indeks kecocokan model menunjukkan bahwa model M2PL lebih mampu memberikan informasi yang lebih banyak dibandingkan dengan model M1PL. Dengan demikian, variasi dua dimensi (kosakata dan verbal) benar-benar lebih mampu menjelaskan data.

Meneruskan hasil analisis pada model pertama, model kedua mengonfirmasi kesetaraan struktur muatan faktor dalam Subtes Verbal berdasarkan kelompok peserta lelaki dan perempuan; apakah dua dimensi (kosakata dan verbal) telah mampu menjelaskan variansi atribut laten yang digunakan para peserta untuk dapat menjawab benar aitem-aitem. Untuk itu dilakukan analisis *multigroup nonlinear confirmatory factor analysis* (MGNCFA) pada kedua kelompok secara simultan. Model awal (MGNCFA-0) adalah model pembatasan parameter aitem sehingga parameter berdasarkan kelompok kedua (lelaki) dibatasi agar sama dengan kelompok pertama (perempuan). Model selanjutnya (MGNCFA-1) adalah model dimana estimasi parameter aitem dibiarkan bebas sehingga mencapai nilai konvergensi.

Tabel 2. Perbandingan Kecocokan Model MGNCFA-0 dan MGNCFA-1

Model	-2loglikelihood	AIC	BIC
MGNCFA-0	416725.04	416847.04	41730 7.39
MGNCFA-1	415710.91	415944.91	41682 7.89
D	1014.13	902.13	479.50

Keterangan: MGNCFA = *multigroup nonlinear confirmatory factor analysis*, MGNCFA-0 = parameter dibuat sama (*constrained*), MGNCFA-1 = parameter dibebaskan (*free*), D = nilai perbedaan kecocokan data, AIC = *Akaike's Information Criterion*, BIC = *Bayesian Information Criterion*.

Hasil yang disajikan pada **Tabel 2** menunjukkan bahwa model bebas (MGNCFA-1) lebih mampu menjelaskan data dibanding model terbatas (MGNCFA-0). Ini didukung oleh kecocokan model yang lebih baik pada model model bebas. Kenyataan ini menunjukkan bahwa estimasi parameter aitem akan lebih baik bila didasarkan pada data masing-masing kelompok, tidak secara bersama-sama. Parameter aitem yang dihasilkan dengan sendirinya menjadi tidak sama persis antara kelompok perempuan dan lelaki. Dengan demikian belum diperoleh model yang mampu menghasilkan kondisi invarians antar kelompok. Dalam kondisi seperti ini, skor kemampuan yang diestimasi berdasarkan model ini menjadi kurang akurat. Bila skor yang diperoleh kelompok perempuan lalu dibandingkan dengan skor pada kelompok lelaki, kedua skor belum sepenuhnya komparabel karena belum tercapai invariansi yang diinginkan.

Model MGNCFA-1 belum secara memuaskan dapat menjelaskan data respons 7000 peserta perempuan dan 7000 peserta lelaki. Peneliti mengajukan model yang lain dengan menambah satu dimensi bersama sehingga terdapat 3 (faktor); yaitu 1 (satu) dimensi bersama, 1 (satu) dimensi spesifik kosakata, dan 1 (satu) dimensi spesifik verbal. Model seperti ini disebut model bifaktor (*bifactor*). Acuan tentang model bifaktor dapat ditelusuri pada beberapa artikel referensi (misalnya Cai dkk., 2011; Liu & Thissen, 2012; Reise, Scheines, Widaman, & Haviland, 2013).

Tabel 3. Perbedaan Kecocokan Model antara MGNCFA-0 dan Bifaktor

Model	-2loglikelihood	AIC	BIC
MGNCFA-1	415710.91	415944.91	416827.89
BIFAKTOR-1	411569.93	411905.93	413173.80
D	4140.98	4038.98	3654.09

Keterangan: MGNCFA = *multigroup nonlinear confirmatory factor analysis*, MGNCFA-1 = parameter dibebaskan (*free*), BIFAKTOR = model bifaktor, D = nilai perbedaan kecocokan data, AIC = *Akaike's Information Criterion*, BIC = *Bayesian Information Criterion*.

Informasi yang disajikan pada **Tabel 3** menunjukkan bahwa model bifaktor 3 (tiga) dimensi (BIFAKTOR-1) memiliki kecocokan data yang lebih baik dibandingkan model 2 (dua) dimensi (MGNCFA-1). Hasil estimasi perbedaan muatan faktor tiap dimensi berdasarkan model bifaktor dituangkan dalam tabel tersendiri.

Tabel 4. Perbedaan Muatan Faktor antara Kelompok Perempuan dan Lelaki Model Bifaktor

No	Aitem	λ_1	λ_2	λ_3
1	i01	-0,06	0,03	-
2	i02	0,04	-0,04	-
3	i04	0,04	-0,05	-
4	i05	0,04	0,04	-
5	i09	0,01	-0,04	-
6	i11	0,02	0,06	-
7	i12	-0,02	0,11	-
8	i13	0,02	0,02	-
9	i15	-0,11	0,09	-
10	i17	-0,05	-	0,02
11	i18	-0,03	-	0,01
12	i20	0,00	-	-0,04
13	i21	0,04	-	0,00
14	i23	-0,03	-	0,00
15	i24	-0,07	-	0,07
16	i25	-0,02	-	0,05
17	i26	-0,08	-	0,02
18	i27	-0,01	-	0,01
19	i29	0,01	-	0,01
20	i30	0,03	-	-0,02
21	i31	0,04	-	0,03
22	i32	-0,02	-	0,00
23	i33	0,02	-	0,06
24	i34	-0,01	-	-0,12
25	i35	-0,05	-	-0,02
26	i36	0,03	-	-0,04
27	i37	0,04	-	-0,04
28	i38	0,03	-	-0,10

Keterangan:
λ_1 = muatan faktor 1 (bersama)
λ_2 = muatan faktor 2 (spesifik)
λ_3 = muatan faktor 3 (spesifik)

Tabel 4 menyajikan perbedaan muatan faktor pada tiap dimensi yang dimodelkan. Pada dimensi bersama (λ_1) hampir tidak ada selisih antara kelompok perempuan dan lelaki, hanya terdapat selisih sebesar -0,11 pada aitem nomor 15 (i15). Aitem nomor 12 (i12) memiliki selisih muatan dimensi kosakata (λ_2) sebesar 0,11. Nilai selisih ini termasuk kecil sehingga dapat diabaikan. Selisih muatan dimensi ketiga (λ_3) sebesar -0,12 pada aitem nomor 34 (i34). Berdasarkan realitas hasil analisis model bifaktor, dapat dikatakan bahwa model ini telah menghasilkan estimasi muatan dimensi yang bersifat invarians antara kelompok perempuan dan lelaki. Dengan demikian kedua kelompok betul-betul berada dalam skala yang sama. Skor-skor hasil estimasi kemampuan pada kelompok perempuan lalu dapat dibandingkan dengan skor-skor pada kelompok lelaki karena kedua kelompok skor sudah sepenuhnya komparabel dan berada dalam skala yang sama.

E. Simpulan, Saran dan Keterbatasan Penelitian

Kesimpulan yang dapat diambil dalam penelitian ini adalah:

1. Korelasi aitem-total (r_{it}) tidak dapat digunakan sebagai bukti validitas skor yang dihasilkan oleh tes;
2. Kesetaraan konstrak dapat dijustifikasi melalui prosedur identifikasi ME/I dengan teknik *multigroup confirmatory factor analysis*, baik yang linier ataupun yang nonlinier;
3. ME/I nonliner merupakan penerapan perbandingan model *multidimensional item response theory* (MIRT) berdasarkan variabel kelompok yang ditetapkan, teknik yang digunakan disebut *multigroup nonlinear confirmatory factor analysis* (MGNCFA);
4. Pemilihan model pengukuran mempengaruhi ME/I antar kelompok.

Rekomendasi yang diberikan berdasarkan hasil studi ini adalah:

1. Hindari penggunaan Korelasi aitem-total (r_{it}) sebagai bukti validitas skor;
2. Laksanakan prosedur ME/I bila penelitian bertujuan membandingkan dua hasil pengukuran;
3. Sebisa mungkin gunakan paradigma model pengukuran *multidimensional item response theory* (MIRT) dalam memodelkan hasil pengukuran.

Beberapa keterbatasan dalam studi ini adalah:

1. Pemilihan model agar model tersebut dapat menghasilkan ME/I antar kelompok bukanlah sesuatu yang mudah. Prosesnya lebih pada proses pencarian (*searching procedure*) yang menggabungkan antara informasi struktur konstrak yang menjadi dasar teoritiknya, sekaligus analisis logis terhadap materi-materi (domain) yang menyusun konstrak tersebut.
2. Prosedur ME/I membutuhkan keterampilan teknis yang rumit bagi peneliti yang belum menguasai psikometri tingkat lanjut sekaligus banyak persoalan pula dalam operasional software yang membantu analisis. Oleh sebab itu penelitian semacam ini terkesan eksklusif bagi peneliti bidang psikologi pada umumnya.

Referensi

APA, AERA, & NCME. (1999). *Standard for educational and psychological testing*. Washington, CD: American Psychological Association.

Bauer, D. J. (2005). The Role of Nonlinear Factor-to-Indicator Relationships in Tests of Measurement Equivalence. *Psychological Methods, 10*(3), 305-316.

Beaujean, A. A., McGlaughlin, S. M., & Margulies, A. S. (2009). Factorial Validity of the Reynolds Intellectual Assessment Scales for Referred Students. *Psychology in the Schools, 46*(10), 932-950.

Booth, T., Irwing, P., & Booth, T. (2011). Sex differences in the 16PF5, test of measurement invariance and mean differences in the US standardisation sample. *Personality & Individual Differences, 50*(5), 553-558.

Bowden, S. C., Saklofske, D. H., & Weiss, L. G. (2011). Intelligence Scale-IV in the United States and Canada Invariance of the Measurement Model Underlying the Wechsler Adult. *Educational and Psychological Measurement, 71*(1), 186–199.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford Press.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*. doi: 10.1037/a0023350

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of

Measurement Invariance. *Structural Equation Modeling, 14*(3), 464–504.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin, 52*, 281-302.

Dimitrov, D. M. (2010). Testing for Factorial Invariance in the Context of Construct Validation. *Measurement and Evaluation in Counseling and Development, 43*(2), 121-149.

Embretson, S. E. (2007). Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure? *Educational Researcher, 36*(8), 449-455.

ETS. (2002). *ETS Standards for Quality and Fairness*. Princeton, NJ: Educational Testing Service.

Flowers, C. P., Raju, N. S., & Oshima, T. C. (2002). *A Comparison Measurement Equivalence Methods Based on Confirmatory Factor Analysis and Item Response Theory*. Paper dipresentasikan pada Annual Meeting of the National Concil on Measurement in Education, New Orleans, LA.

Gorin, J. S. (2007). Reconsidering Issues in Validity Theory. *Educational Researcher, 36*(8), 456–462.

Immekus, J. C., & Maller, S. J. (2010). Factor Structure Invariance of the Kaufman Adolescent and Adult Intelligence Test across Male and Female Samples. *Educational and Psychological Measurement, 70*(1), 91-104.

Jones-Farmer, L. A. (2010). The Effect of Among-Group Dependence on the Invariance Likelihood Ratio Test. *Structural Equation Modeling, 17*(3), 464–480.

Jöreskog, K. G. (2007). Factor Analysis and Its Extensions. Dalam R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: historical developments and future directions* (hh. 47-77). Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Liu, Y., & Thissen, D. (2012). Identifying Local Dependence With a Score Test Statistic Based on the Bifactor Logistic Model. *Applied Psychological Measurement, 36*(8), 670-688. doi: 10.1177/0146621612458174

Maydeu-Olivares, A., Hernández, A., & McDonald, R. P. (2006). A Multidimensional Ideal Point Item Response Theory Model for Binary Data. *Multivariate Behavioral Research, 41*(4), 445–471.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

McDonald, R. P. (2000). A Basis for Multidimensional Item Response Theory. *Applied Psychological Measurement, 24*(2), 99–114.

Molenaar, P. C. M. (2009). Commentary on "Idiographic Filters for Psychological Constructs". *Measurement, 7*(1), 13-16.

Naga, D. S. (2004). Ketidaktepatan Penggunaan Validitas Butir dan Koefisien Reliabilitas dalam Penelitian Pendidikan dan Psikologi. *Jurnal Ilmu Pendidikan, 11*(2).

Nair, R. L., White, R. M. B., Knight, G. P., & Roosa, M. W. (2009). Cross-Language Measurement Equivalence of Parenting Measures for Use With Mexican American Populations. *Journal of Family Psychology, 23*(5), 680–689.

Ployhart, R. E., & Oswald, F. L. (2004). Applications of Mean and Covariance Structure Analysis: Integrating Correlational and Experimental Approaches. *Organizational Research Methods, 7*(1), 27-65.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor

Perspective. *Educational and Psychological Measurement, 73*(1), 5-26. doi: 10.1177/0013164412449831

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance. *Psychological Bulletin, 114*(3), 552-566.

Reniers, R. L. E. P., Corcoran, R., Drake, R., Shryane, N. M., & Völlm, B. A. (2011). The QCAE: A Questionnaire of Cognitive and Affective Empathy. *Journal of Personality Assessment, 93*(1), 84-95.

Rotgans, J., & Schmidt, H. (2009). Examination of the Context-Specific Nature of Self-Regulated Learning. *Educational Studies, 35*(3), 239-253.

Sireci, S. G. (2007). On Validity Theory and Test Validation. *Educational Researcher, 36*(8), 477–481.

South, S. C., Krueger, R. F., & Iacono, W. G. (2009). Factorial Invariance of the Dyadic Adjustment Scale across Gender. *Psychological Assessment, 21*(4), 622-628.

Vandenberg, R. J. (2002). Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures. *Organizational Research Methods, 5*(2), 139-158.

Whittaker, T. A., Chang, W., & Dodd, B. G. (2012). The Performance of IRT Model Selection Methods With Mixed-Format Tests. *Applied Psychological Measurement, 36*(3), 159-180. doi: 10.1177/0146621612440305

Yao, L. (2010). BMIRT: Bayesian multivariate item response theory. [Computer Software]. Monterey, CA: Defense Manpower Data Center.

Yao, L., & Li, F. (2010). *A DIF Detection Procedure in Multidimensional Item Response Theory Framework and its Applications*. Paper dipresentasikan pada Annual Meeting of the National Council on Measurement in Education, Colorado, Denver.