ROBUSTNESS MODEL-MODEL RESPONS BUTIR TERHADAP PELANGGARAN ASUMSI INDEPENDENSI LOKAL BUTIR

[1)]Ali Hasmy, [2)]Suryanto, [3)]Kumaidi
[1)]Sekolah Tinggi Agama Islam Negeri Pontianak,
[2)]Universitas Negeri Yogyakarta,
[3)]Universitas Muhammadiyah Surakarta
[1)]ali_hasmy@yahoo.com, [2)]suryanto@uny.ac.id[3)]kuma_426@yahoo.com

Abstrak

Tujuan utama penelitian ini adalah untuk mengetahui robustness Model Logistik 1 Parameter (ML 1-P), Model Logistik 2 Parameter (ML 2-P) dan Model Logistik 3 Parameter (ML 3-P) terhadap pelanggaran Asumsi Independensi Lokal Butir (ILB). Penelitian ini menggunakan data simulasi yang dibangkitkan dengan 40 butir, 500 simuli, dan 10 replikasi untuk setiap model. Skor-skor batas dibangun berdasakan pelanggaran Asumsi ILB 0 - 100% yang dihasilkan dengan menggunakan 1- 40 kelompok butir sedangkan kategori-kategori skor dibangun berdasarkan dampak pelanggaran Asumsi ILB terhadap struktur dari matriks data. Hasil penelitian ini menunjukkan bahwa model yang paling robust terhadap pelanggaran Asumsi ILB adalah ML 1-P dengan skor batas 31,71% (kategori pelanggaran berat) diikuti ML 2-P dengan skor batas 12,1% (kategori pelanggaran sedang), dan ML 3-P dengan skor batas 7,68% (kategori pelanggaran sedang).

Kata kunci: *model-model respons butir, robustness, independensi lokal butir*

ROBUSTNESS MODEL-MODEL RESPONS BUTIR TERHADAP PELANGGARAN ASUMSI INDEPENDENSI LOKAL BUTIR

[1)]Ali Hasmy, [2)]Suryanto, [3)]Kumaidi
[1)]Sekolah Tinggi Agama Islam Negeri Pontianak,
[2)]Universitas Negeri Yogyakarta,
[3)]Universitas Muhammadiyah Surakarta
[1)]ali_hasmy@yahoo.com, [2)]suryanto@uny.ac.id[3)]kuma_426@yahoo.com

Abstract

The primary purpose of this study was to investigate the robustness of the 1-PLM, 2-PLM, and 3-PLM, against violation of the Local Item Independence (LII) Assumption based on cut-off scores and score categories. The investigation used simulated data generated with 40 items, 500 simulees, and 10 replications for each model. The cut-off scores were built based-on 0 – 100% violations of the LII Assumption that were introduced using 1- 40 item clusters. The score categories in this study were built based-on impact of the violations of the LII Assumption to the structure of data matrix. The result showed that the most robust model was 1-PLM with cut-off score 31,71% (heavy violation category) followed by 2-PLM with cut-off score 12,1% (moderate violation category), and 3-PLM with cut-off score 7,68% (moderate violation category).

Keywords: *item response models, robustness, local item independence*

Pendahuluan

Menurut Embretson (1996, p. 341), Sijtsma & Junker (2006, pp.75 & 77), dalam psikometrika, C*lassical Test Theory* merupakan pendekatan statistika yang dominan digunakan sampai Lord & Novick pada tahun 1968 merilis buku mereka "*Statistical theories of mental test scores*". Karya Lord & Novick dengan tambahan kontribusi dari Birnbaum ini dalam beberapa hal menurut Embretson & Reise (2000, p.5) merupakan perluasan dari eksposisi Teori Tes Klasik (TTK) Gullicksen yang dituangkan dalam bukunya "*Theory of mental tests*" tahun 1950. Meskipun demikian, *milestone* dari *Item Response Theory* sebenarnya adalah karya Lord (1952) yaitu "*A theory of test score*".

Jika ditelusuri lebih jauh, cikal bakal Teori Respons Butir (TRB) pada dasarnya telah dirintis oleh Thurstone sejak tahun 1925 (Wright & Stone, 1979, p.vii) dengan karya-karyanya, khususnya tentang penskalaan *Case V* (Andrich dalam Gorard, 2008, pp.66-78). Penskalaan ini merupakan basis bagi *Item Response Models* khususnya *Parametric Item Response Models* yang bersifat Probabilistik (*Probabilistic Item Response Models*).

Model-model Respons Butir (MRB) Probabilistik Parametrik yang banyak digunakan adalah model-model unidimensional untuk skor dikotomus. Model-model dimaksud menurut Johnson (2007, pp.4-5) adalah ML 1-P atau Model Rasch, ML 2-P dan ML 3-P dari Birnbaum. Tiga MRB yang popular inilah yang dikaji pada penelitian ini.

Menurut Hulin, Drasgow, & Parsons (1983, p.38) serta Wainer, Bradlow, & Wang (2007, p.25) ML 1-P dapat dituliskan sebagai berikut:

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

Dengan e = 2,718… adalah basis logaritma natural, θ_i adalah parameter abilitas peserta tes, b_j adalah *threshold* yang berhubungan dengan kesulitan dari butir ke-*j*. Semakin besar nilai b_j berarti butir yang bersangkutan semakin sulit dan sebaliknya. ML 2-P dari Birnbaum menurut Hulin, Drasgow, & Parsons (1983, p.36), Torgerson (1958, pp.202-203), dan du Toit (2003, p.539) adalah:

$$P_i(\theta) = \frac{1}{1+e^{-a_j(\theta_i - b_j)}}$$

dengan a_j adalah *slope* butir yang berhubungan dengan daya pembeda butir ke-*j*. Sedangkan ML 3-P dari Birnbaum menurut Fox (2010, p.11) serta Wainer, Bradlow, & Wang (2007, p.30) adalah:

$$P_i(\theta) = c_j + (1 - c_j)\frac{1}{1+e^{-a_j(\theta_i - b_j)}}$$

dengan c_j adalah *pseudo-chance level* pada butir ke-*j*.

TRB ini menurut Ip, Wang, De Boeck, & Meulders (2004, p.191) telah terbukti merupakan alat yang sangat ampuh dalam memodelkan respons individu terhadap stimuli behavioral tertentu. Pada TRB diterapkan *Conjoint Measurement* karena, menurut Bond & Fox (2007, pp.8-11,) pada model-model TRB khususnya yang termasuk dalam kategori *General/Common Item Response Models*, probabilitas *examinee* menjawab benar berubah sesuai dengan nilai dua atribut yang lain yaitu *person ability* dan *item difficulty*. Istilah *item difficulty* ini oleh Brown (2005, p.66) disebut dengan *item facility*.

Tidak seperti TTK, TRB khususnya Model Rasch memenuhi *specific objectivity* atau memenuhi sifat "*person distribution free*" karena memperhitungkan abilitas peserta tes pada estimasi tingkat kesulitan butir dan memenuhi sifat "*item distribution free*" karena memperhitungkan kesulitan butir pada estimasi abilitas peserta tes. *Spesific objectivity* ini membawa implikasi dapat dilakukannya *parameter separation* yang nilai suatu parameter dapat diestimasi tanpa mengetahui nilai-nilai dari parameter yang lain. Dengan demikian, maka identitas suatu variabel dari satu peristiwa ke peristiwa yang lainnya

menjadi konstan (*invariance*). Sifat ini merupakan persyaratan *prima facie* dan merupakan sifat yang krusial agar pengukuran secara ilmiah dapat diwujudkan. Selain itu, sifat ini juga terkait dengan upaya formulasi dari kebutuhan terhadap *generality* dan *validity* dari suatu pengukuran.

TRB juga memiliki beberapa kelebihan lain dibandingkan dengan TTK antara lain bahwa, menurut TRB, tes yang pendek dapat sama reliabelnya dengan tes yang panjang, lebih mudah menangani format campuran (Embretson, 1996, p.342; Panther & Reeve, 2002, p.S24), dan skor yang dihasilkan lebih merupakan variabel daripada skor yang dijumlah pada TTK (Edwards, 2009, p.519). Selain itu, kekurangan utama TTK adalah mengasumsikan bahwa presisi pengukuran bersifat konstan antarkeseluruhan rentang sifat yang dikaji pada semua peserta tes. Cara penskoran dengan menggunakan rerata maupun penjumlahan pada TTK juga dapat membawa pada kekeliruan jika digunakan untuk inferensi (Fraley, Waller, & Brennan, 2000, p.353).

Berbagai kelebihan yang ada pada TRB sebagaimana yang telah dipaparkan di atas membuat, sejak karya para pionir TRB seperti buku dari Lord & Novick yang fundamental serta buku dari Fischer "*Introduction to the theory of psychological tests*" pada tahun 1974, TRB telah berkembang pesat di berbagai bidang seperti psikologi dan pendidikan. Bahkan, pada saat sekarang ini TRB telah mendominasi literatur kependidikan di beberapa negara dan secara perlahan bergerak ke dalam aplikasi ilmu sosial lainnya. Dominasi ini telah berlangsung selama dua dekade terakhir melalui penggunaan TRB untuk memeriksa *item adequacy checking*, *test scoring*, *score equating*, dan pembuatan *Computer Adaptive Test* (CAT).

TRB yang berbasiskan variabel laten ini, merupakan dasar bagi teoretisasi domain seperti *output* kesehatan, *output* pendidikan, risiko, kepercayaan, sindrom klinis, maupun kepribadian (Johnson, 2007, pp.1-2, Panther & Reeve, 2002, pp.S21-S22). Hal ini dikarenakan dalam *behavioral sciences*, banyak konstruks, baik secara teoretik maupun praktik, tidak dapat diamati secara langsung. Prosedur dasar untuk mengukur suatu konstruks yang sedemikian rupa adalah dengan melibatkan *observed variables*, yang dapat memberikan bukti tak langsung bagi konstruks yang diteliti (*inference*). Biasanya, hal ini berarti bahwa suatu tes dikembangkan dengan menyusun beberapa indikator dan butir-butir yang merepresentasikan konstruks yang diteliti. Pada saat ini, acuan yang membumi dan lebih mendukung tipe "Inferensi" seperti ini adalah TRB, sedangkan TTK pada dasarnya lebih bersifat "Deskriptif" (Bradlow, Wainer, & Wang, 1998, p.1; Braeken & Tuerlinckx, 2009, p.1127).

Sifat inferensi yang melekat pada TRB membawa konsekuensi, terutama pada model-modelnya yang secara umum merupakan model parametrik, yaitu perlunya dilakukan *Model Adequacy Checking*. Oleh karena itulah Hambleton & Jones (1993, p.258) menyatakan bahwa TRB didasari oleh asumsi-asumsi yang kuat, tidak seperti TTK yang lebih bersifat deskriptif. Namun, yang perlu diingat bahwa asumsi-asumsi yang kuat sebenarnya lebih dialamatkan pada MRB Parametrik yang lebih bersifat probabilistik dibandingkan dengan MRB Nonparametrik yang lebih bersifat deterministik dengan berdasarkan asumsi-asumsi yang minimum.

Sehubungan dengan asumsi pada penerapan MRB Parametrik, termasuk 1-PLM, 2-PLM, dan 3-PLM, sebagaimana dimaksudkan di atas, ada beberapa asumsi yang perlu diperhatikan, antara lain sebagai berikut: (1) Probabilitas θ mengikuti distribusi tertentu yang umumnya adalah normal (Hulin, Drasgow, & Parsons, 1983, p.39; Ramsay, 1997, p.383; Traub, 1983, p.58). (2) Derajat pengukuran θ diasumsikan interval atau rasio (Mokken, 1997, p.351; van der Linden & Hambleton, 1997, p.348). (3) Monotonisitas hubungan antara sifat laten dan performansi pada setiap butir tes (Mokken, 1997, pp.353-354, pp.374-379). (4) *Nonspeeded test administration* (Hambleton & Swaminathan, 1985, p.30; Hambleton, Swaminathan, & Rogers, 1991, p.57). (5) Pengukuran

bersifat unidimensi (Hambleton & Swaminathan, 1985, pp.16-22; Hulin, Drasgow, & Parsons, 1983, p.40; Spray, 1997, p.210; Traub, 1983, p.58); (6) Independensi lokal (Hambleton & Swaminathan, 1985, pp.22-25; Hulin, Drasgow, & Parsons, 1983, p.41; Spray, 1997, p.211).

Di antara beberapa asumsi MRB Parametrik, yang paling sering dibicarakan adalah Asumsi Unidimensionalitas dan Asumsi Independensi Lokal (IL). Karenanya kedua asumsi ini disebut oleh Hambleton, Swaminathan, & Rogers (1991, p.9) sebagai *common assumptions* dalam penerapan MRB Parametrik.

Hubungan antara kedua asumsi ini sangat erat dan sering dianggap paralel. Namun, berdasarkan pendapat Hambleton & Swaminathan (1985, pp.22 & 24) yang menyatakan bahwa Asumsi Independensi Lokal Butir terpenuhi (dependensi = 0%) bila θ bersifat unidimensi dan pendapat Crocker & Algina (1986, p.343) yang menyatakan bahwa unidimensionalitas akan tercapai bila seluruh butir dependen secara statistik (dependensi=100%), jelas memperlihatkan pola hubungan yang tidak paralel.

Asumsi IL sebagaimana dimaksudkan di atas dapat dinyatakan sebagai berikut:

$$P\left(Y_{ij}=y_{ij}|\theta_i\right)=\prod_{i=1}^{I}\prod_{j=1}^{J}P\left(y_{ij}|\theta_i\right)\cdot$$

Independensi ini mencakup Independensi Lokal Person (ILP) dan Independensi Lokal Butir (ILB). ILP disebut juga Independensi Lokal Interperson atau *Local/Conditional Independence between Persons* yang dapat dituliskan sebagai berikut:

$$P\left(Y_j=y_j|\theta_i\right)=\prod_{i=1}^{I}P\left(y_{ij}|\theta_i\right)=P\left(y_{1j}|\theta_i\right)P\left(y_{2j}|\theta_i\right)...P\left(y_{Ij}|\theta_i\right)$$

ILB disebut juga Indepen-densi Lokal Intraperson atau *Local/Conditional Independence within Person* yang dapat dituliskan sebagai berikut:

$$P\left(Y_i=y_i|\theta_i\right)=\prod_{j=1}^{J}P\left(y_{ij}|\theta_i\right)=P\left(y_{i1}|\theta_i\right)P\left(y_{i2}|\theta_i\right)...P\left(y_{iJ}|\theta_i\right)$$

Pada penelitian ini Asumsi IL yang dikaji adalah Asumsi ILB. Oleh karena itu, maka uraian selanjutnya dikhususkan pada asumsi ini.

Berkaitan dengan Asumsi ILB ini, maka perlu diingat bahwa pada suatu tes unidimensi probabilitas sukses peserta tes pada butir tertentu hanya tergantung pada parameter butir yang bersangkutan serta pada abilitas peserta tes (θ), tidak pada hal lain. Jika MRB yang digunakan *fit with data*, maka abilitas peserta tes (θ) merupakan satu-satunya faktor yang menentukan kesuksesan atau ketidaksuksesan peserta tes pada butir dimaksud. Menurut Hulin, Drasgow & Parsons (1983, p.42), hal ini berarti bahwa, pengetahuan peserta tes pada butir yang lain tidak memberikan tambahan informasi apapun pada penentuan kesuksesan atau ketidaksuksesan tersebut. Jika pengetahuan peserta tes pada butir yang lain memberikan tambahan informasi, maka menurut Lord (1980, p.19), performansi peserta tes pada butir yang bersangkutan sudah tentu tergantung pada hal lain selain θ yang telah dispesifikasikan. Hal ini tentu saja bertentangan dengan asumsi dari Lazarsfeld mengenai IL.

Sebelumnya, Lord (1952, p.8) juga menyatakan bahwa antarpasangan butir haruslah independen sebagai konsekuensi dari pembatasan mengenai *common factor*. Hal ini dikarenakan bila ada pasangan butir yang saling dependen berarti ada *common factor* yang menghubungkan keduanya. Selain itu, pernyataan Lord ini juga merupakan suatu konsekuensi logis, karena MRB merupakan salah satu kelas dalam *Latent Structure Analysis* yang menurut Anderson (1959, p.1) memiliki salah satu "Asumsi Utama" yaitu, respons pada butir yang berbeda adalah independen.

Pelanggaran terhadap Asumsi ILB tersebut di atas, dapat mengakibatkan estimasi dan reliabilitas dari parameter-parameter model diragukan (Braeken, Tuerlinckx, & De Boeck, 2005, p.2). Hal ini sejalan dengan pendapat Bradlow, Wainer, & Wang (1998, p.3) yang menyatakan bahwa akibat pelanggaran Asumsi IL dapat terjadinya pernyataan yang berlebihan dari presisi estimasi abilitas peserta tes, bias pada estimasi

tingkat kesulitan dan daya pembeda butir. Akibat pelanggaran Asumsi ILB terhadap estimasi abilitas peserta tes diperjelas oleh Scott & Ip (2002, p. 2) yang menyatakan bahwa bias lebih besar untuk sub-sub-kelompok dengan rata-rata yang lebih besar. Bias pada estimasi parameter daya pembeda butir dipertegas oleh Orlando (2008, p.3), yang menyatakan bahwa pelanggaran Asumsi ILB mengakibatkan terjadinya inflasi pada estimasi *slope* (parameter daya pembeda). Bias pada estimasi parameter model, baik parameter abilitas peserta tes maupun parameter-parameter butir, menurut Antal (2003, p.4) dapat mengakibatkan model tidak sesuai dengan data (*misfit*). Selain itu, menurut Kim, Ayala, Ferdous, & Nering (2007, p.17) pelanggaran Asumsi ILB dapat mengakibatkan *overestimate* pada informasi butir dan informasi tes.

Dampak ikutan dari hal-hal tersebut di atas menurut Goodman & Luecht (2009, p.6) adalah sebagaimana yang dicantumkan berikut ini: (1) Jika residual kovariansi berbeda dari berbagai subkelompok populasi, maka akan terjadi *differential item functioning*; (2) Pembuatan bank butir (*item bank*) dapat diragukan karena ketidakakuratan pada estimasi parameter butir; (3) Penyusunan tes paralel dari bank butir juga diragukan; (4) Penskalaan dan penyetaraan juga dapat ikut diragukan.

Jauh sebelumnya, Lord & Novick (1968, p.436) menyatakan bahwa Asumsi ILB terkait dengan sifat *sufficiency* dan efisiensi secara statistik dari *estimator* serta dalam hal *classification rule*. Oleh karena itulah, maka Hambleton, Swaminathan, & Rogers (1991, pp.10-12) menguraikan Asumsi ILB, selain Asumsi Unidimensionalitas, secara khusus. Bahkan, Scott & Ip (2002, p.2) dengan tegas menyebut asumsi ini sebagai "*basic premise*". Hal ini sejalan dengan pendapat Zwinderman (1997, p.248) yang menyatakan bahwa "*most important assumption*" dalam penerapan MRB adalah $\sum X_i$ merupakan *sufficiency statistic* minimal bagi θ_i, maka respons-respons pada butir tes haruslah independen secara lokal untuk θ tertentu.

Mengingat banyaknya persoalan yang dapat muncul berkaitan dengan pelanggaran Asumsi ILB, maka berbagai penelitian telah dilakukan. Penelitian-penelitian dimaksud antara lain dilakukan oleh oleh Yen pada tahun 1984, Ackerman pada tahun 1987, Reese pada tahun 1995, Huynh, Michels, & Ferrara pada tahun 1995, Bradlow, Wainer, & Wang pada tahun 1998, Tuerlinckx & De Boeck pada tahun 2001, Scott & Ip pada tahun 2002, Stark, Chernyshenko, & Drasgow pada tahun 2002, Jiao & Kamata pada tahun 2003, Zenisky, Hambleton, & Sireci pada tahun 2003, Wang & Wilson pada tahun 2005, Kim, Ayala, Ferdous, et al. pada tahun 2007, Pommerich & Ito pada tahun 2008, Mislevy pada tahun 2011, serta Jiao, Kamata, Wang, et al. pada tahun 2012. Penelitian-penelitian ini umumnya terfokus pada tiga hal yaitu: (a) Metode untuk melakukan pemeriksaan terhadap pelanggaran Asumsi ILB; (b) Efek negatif jika pelanggaran tersebut terjadi; dan (c) Pembangunan Model (*Model Building*) Respons Butir yang dapat mengatasi efek negatif dimaksud (*robust* terhadap pelanggaran asumsi ILB).

Namun, karena unidimensionalitas secara utuh sulit dipenuhi maka pelanggaran Asumsi ILB hampir selalu ada. Konsekuensinya, model respons butir khusus untuk mengatasi efek negatif DLB menjadi begitu superior. Karenanya, perlu diteliti mengenai batas (*cut-off*) proporsi pelanggaran Asumsi ILB yang dapat ditoleransi sehingga penerapan ML 1-P, ML 2-P, maupun ML 3-P sebagai MRB yang banyak digunakan dalam praktik (*popular IRMs*) tetap layak untuk dilakukan. Pada saat yang sama, dengan mengetahui batas dimaksud, dapat diketahui juga model mana yang lebih *robust* terhadap pelanggaran Asumsi ILB.

Metode Penelitian

Penelitian ini merupakan Studi Monte Carlo (SMC) dengan menggunakan desain faktorial sebagaimana terlihat pada model

berikut ini yang mengacu pada Fox (2008, p. 163).

$$\eta_i = \beta_0 + \beta_1 Y_{G_1} + \beta_2 Y_{G_2} + \beta_3 Y_{G_3} + \beta_4 Y_{C_1} + \beta_5 Y_{C_2} + \beta_6 Y_{C_3} + \beta_7 Y_{G_1} Y_{C_1} + \beta_8 Y_{G_1} Y_{C_2} + \beta_9 Y_{G_1} Y_{C_3} + \beta_{10} Y_{G_2} Y_{C_1} + \beta_{11} Y_{G_2} Y_{C_2} + \beta_{12} Y_{G_2} Y_{C_3} + \beta_{13} Y_{G_3} Y_{C_1} + \beta_{14} Y_{G_3} Y_{C_2} + \beta_{15} Y_{G_3} Y_{C_3}$$

Penelitian ini dilaksanakan dalam kurun waktu kurang lebih selama satu tahun mulai Agustus 2012 sampai dengan April 2013. Tempat penelitian ini umumnya adalah di Laboratorium Komputer Pascasarjana serta Laboratorium Penelitian dan Evaluasi Pendidikan (Perpustakaan Pascasarjana) Universitas Negeri Yogyakarta (UNY).

Data yang digunakan pada penelitian ini bersumber dari hasil simulasi yang dilakukan oleh peneliti sendiri dengan menggunakan perangkat lunak *MS Excel.*

Variabel penelitian ini yakni:

1. Variabel Bebas yaitu: (a) MRB yang digunakan untuk melakukan membangkitkan data yaitu ML 1-P, ML 2-P, dan ML 3-P untuk data dikotomus, dan (b) MRB yang digunakan untuk melakukan kalibrasi data bangkitan yang juga menggunakan ML 1-P, ML 2-P, dan ML 3-P untuk data dikotomus.
2. Variabel Terikat yaitu rerata persentase butir yang *fit* dengan menggunakan kriteria *Chi-Square* pada pengelompokan butir sebanyak 1 – 40 (sebanyak butir).

Berdasarkan poin 1 dan 2, maka desain penelitian ini adalah desain faktorial 3 x 3 x 40 atau dengan kasus sebanyak 360.

Pada SMC ini, data dikumpulkan dengan membangkitkannya menggunakan perangkat lunak *MS Excel* 2010 dengan langkah-langkah: (a) membangkitkan *True θ's* untuk 500 simuli dengan menggunakan distribusi normal baku; (b) membangkitkan *True Item Parameters* untuk 40 butir dengan menggunakan distribusi normal dimana parameter tingkat kesulitan menggunakan rentang -2,0 – 2,0 (Hambleton & Swaminathan, 1985,p. 107), parameter daya beda menggunakan rentang 0,0 – 2,0 (Hambleton & Swaminathan, 1985, p.36) dan parameter *pseudo chance level* menggunakan rentang 0 – 0,25 dengan asumsi ada 4 pilihan jawaban (Hullin, Drasgow, & Parsons, 1983, p.36); (c) membangkitkan *Response Data Sheets* dengan cara menghitung $P_{ij}(\theta)$, membangkitkan bilangan random $U(0,1)$, serta membangkitkan simulasi nilai amatan Y dengan membandingkan nilai $P_{ij}(\theta)$ dengan $U(0,1)$ pada sel yang bersesuaian; dan (d) melakukan replikasi sebanyak 10 kali untuk tiap-tiap MRB.

Analisis data dilakukan dengan langkah-langkah sebagai berikut. **Pertama**, Memeriksa pelanggaran Asumsi ILB pada 780 pasangan butir (*pairwise*). Banyaknya pasangan butir ini didapatkan melalui:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Dengan n adalah banyaknya butir yaitu 40, x adalah 2 butir yang dipasangkan (Freund's, 2004, p.8).

Dengan demikian, untuk 1.200 *sheet* data (3 model x 40 kelompok butir x 10 replikasi) diperlukan 936.000 kali analisis dengan menggunakan prosedur Q_3 dari Yen sebagai berikut:

$$Q_3 = r_{d_j d_{j'}}$$

Dengan r adalah korelasi *product moment* dari Pearson antara d_j dan $d_{j'}$, $d_j$adalah residual butir ke-j, $d_{j'}$ adalah residual butir ke-j' (Balazs & De Boeck, 2007 p.12; Goodman & Luecht, 2009, p.11; Reese, 2005 p.2).

Kedua, menghitung persentase pelanggaran Asumsi ILB dengan formula sebagai berikut:

$$\%LID = \frac{number\ \ of\ \ dependence\ \ item\ \ pairs}{number\ \ of\ \ item\ \ pairs} \times 100$$

Persentase ini kemudian dihitung reratanya dari 10 replikasi untuk masing-masing kelompok butir pada setiap MRB yang digunakan untuk membangkitkan data.

Ketiga, menentukan skor batas (*cut-score*) *robustness* MRB yang dilakukan dengan cara: (a) menggunakan nilai signifikansi 5% sebagai batas *robust* atau tidaknya penggunaan MRB pada tiap-tiap butir; (b) menghitung rerata persentase butir yang *fit* pada tiap-tiap data *sheet* dan menghitung rerata

gabungannya untuk 10 replikasi; (c) menggunakan batas bahwa estimasi interval dengan signifikansi 5% untuk rerata populasi sebagai *cut-score* dari *robustness* MRB secara keseluruhan.

Keempat, membuat kategori-kategori skor pelanggaran Asumsi ILB berdasarkan dampaknya terhadap struktur dari matriks data dan

Kelima, membandingkan MRB dengan cara deskriptif berdasarkan *cut-off scores* dan *score categories* pelanggaran Asumsi ILB. Selain itu perbandingan juga dilakukan secara inferensia dengan menggunakan analisis Model Linear yang Digeneralisasi (*Generalized Linear Model*) dengan dua faktor. Model ini menurut Fox (2008, p.379) memiliki tiga komponen berikut ini.

1. Suatu komponen random dari variabel respons Y yang mengikuti keluarga distribusi eksponensial yang dalam penelitian ini adalah distribusi binomial yaitu:

$$p(y)=\binom{n}{ny}\mu^{ny}(1-\mu)^{n(1-y)}.$$

Dalam hal ini nx adalah banyaknya observasi sukses (skor 1) dalam n percobaan dan $n(1-x)$ banyaknya ketidaksuksesan.

Koefisien binomialnya diberikan oleh:

$$\binom{n}{ny}=\frac{1!}{(ny)!\{n(1-y)\}!}.$$

2. Suatu prediktor linear yang merupakan suatu fungsi linear dari regresi, yang pada penelitian ini adalah:

$$\eta_i=\beta_0+\beta_1Y_G+\beta_2Y_C+\beta_3Y_GY_C.$$

3. Suatu *link function* g (.) yang pada penelitian ini adalah *logit link* yang merupakan *canonical link* dari distribusi binomial, yaitu:

$$\eta_i=g(\mu_i)=\log_e\frac{\mu_i}{1-\mu_i}$$

dengan inversnya,

$$\mu_i=g^{-1}(\eta_i)=\frac{1}{1+e^{-\eta_i}}$$

yang mentransformasi ekspektasi dari variabel respons ke prediktor linear yaitu:

$$g(\mu_i)=\eta_i=\beta_0+\beta_1Y_G+\beta_2Y_C+\beta_3Y_GY_C.$$

Karena *link function* bersifat *invertible* maka dapat juga ditulis,

$$\mu_i=g^{-1}(\eta_i)=g^{-1}(\beta_0+\beta_1Y_G+\beta_2Y_C+\beta_3Y_GY_C)$$

Untuk menguji model yang dispesifikasikan (model 1) dibandingkan dengan model 0, maka digunakan *likelihood-ratio test statistic*:

$$G_0^2=D_0-D_1$$

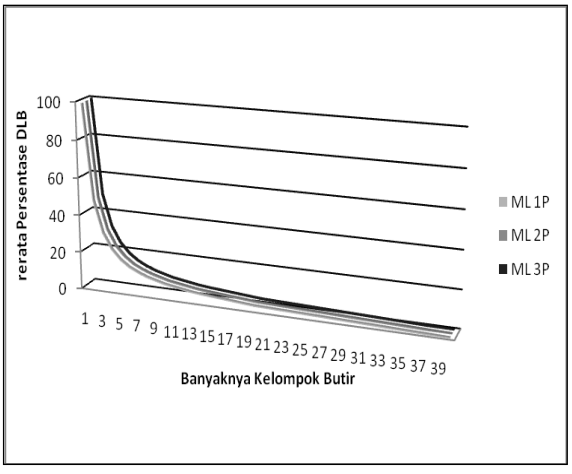
yang mengikuti distribusi *chi-square* dengan derajat kebebasan $k_1 - k_0$. Sedangkan untuk melakukan uji parameter-parameternya digunakan *Wald Statistic*:

$$Z_0=\left(B_j-\beta_j^{(0)}\right)/SE\left(B_j\right),$$

Dengan $SE\left(B_j\right)$ adalah *asymptotic standard error* dari koefisien B_j yang diestimasi.

Temuan dan Diskusi

Hubungan antara banyaknya kelompok butir dan rerata persentase DLB untuk data yang dibangkitkan dengan menggunakan ML 1-P, ML 2-P, dan ML 3-P dapat dilihat pada Gambar 1.



Gambar 1. Hubungan antara Banyaknya Kelompok Butir dan Rerata Persentase Pelanggaran Asumsi ILB

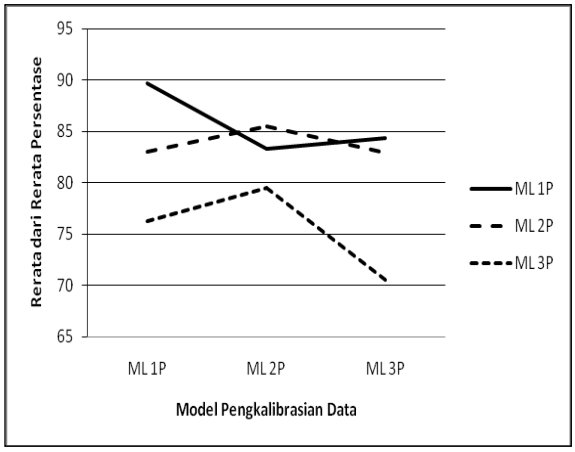
Gambar 1 memperlihatkan bahwa hubungan banyaknya kelompok butir dengan rerata persentase pelanggaran Asumsi ILB pada data yang dibangkitkan dengan ML 1-P, ML 2-P, maupun ML 3-P jelas memperlihatkan pola yang sama (konsisten). Dalam bentuk regresi, model hubungan banyaknya pengelompokan butir dengan rerata persentase pelanggaran Asumsi ILB mengikuti model regresi *inverse*, dengan persamaan untuk ML 1-P, ML 2-P, dan ML 3-P secara berurutan adalah:

$E(Y_t) = -2{,}313 + 101{,}174/t$
$E(Y_t) = -2{,}337 + 101{,}347/t$
$E(Y_t) = -2{,}322 + 101{,}924/t.$

Kesemua persamaan regresi yang tercantum di atas memiliki *p-value* < 0,001. Hal ini memberikan bukti bahwa semakin sedikit pengelompokan butir cenderung semakin besar rerata pelanggaran Asumsi ILB, sebaliknya jika pengelompokan butir semakin banyak (mendekati banyaknya butir), maka rerata pelanggaran Asumsi ILB semakin kecil. Selain itu, jika mengacu pada pendapat Crocker & Algina (1986, p.343) maka unidimensionalitas terjadi jika seluruh butir mengelompok menjadi satu, sebaliknya jika mengacu pada pendapat Hambleton & Swaminathan (1985, pp.22 & 24) maka unidimensionalitas terjadi jika banyaknya kelompok butir sama dengan banyaknya butir.

Sementara itu hubungan antara MRB yang digunakan dengan rerata persentase butir yang *fit* dapat dilihat pada Gambar 2 dan Gambar 3 tentang rerata dan simpangan baku dari rerata persentase butir yang *fit* untuk data yang dibangkitkan dan dikalibrasi dengan ML 1-P, ML 2-P, dan ML 3-P.

Gambar 2 memperlihatkan bahwa pengkalibrasian dengan ML 3-P untuk data yang dibangkitkan dengan ML 1-P, ML 2-P, maupun ML 3P, ternyata menghasilkan rerata persentase butir yang *fit* paling sedikit jika terdapat pelanggaran Asumsi ILB. Pengkalibrasian dengan ML 1-P untuk data yang dibangkitkan dengan ML 2-P dan ML 3-P menghasilkan rerata persentase butir yang *fit* paling banyak jika terdapat pelanggaran Asumsi ILB, tetapi tidak jika data dibangkitkan dengan ML 2-P. Sedangkan pengkalibrasian dengan ML 2-P menghasilkan rerata persentase butir yang *fit* paling banyak, jika terdapat pelanggaran Asumsi ILB, untuk data yang dibangkitkan dengan ML 2-P saja. Tetapi secara umum, rerata persentase butir yang *fit* yang dihasilkan dengan menggunakan ML 1-P untuk kalibrasi adalah paling stabil hasilnya dibandingkan dengan pengkalibrasian menggunakan ML 2-P dan ML 3-P.

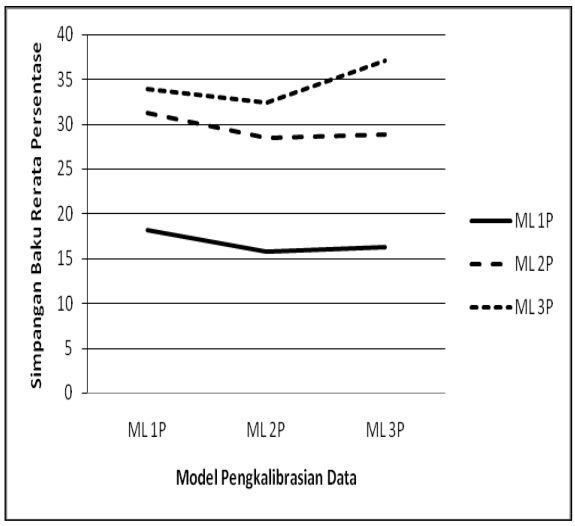


Gambar 2. Rerata Persentase Butir yang *Fit* untuk Data yang Dibangkitan dengan ML 1-P, ML 2-P, dan ML 3-P

Sementara itu Gambar 3 memperlihatkan bahwa tidak peduli data dibangkitkan dengan ML 1-P, ML 2-P, maupun ML 3-P, pengkalibrasian dengan menggunakan ML 1-P memberikan Simpangan Baku (SB) yang selalu lebih kecil. Sebaliknya, tidak peduli data dibangkitkan dengan ML 1-P, ML 2-P, maupun ML 3-P, pengkalibrasian dengan menggunakan ML 3-P selalu menghasilkan SB yang lebih besar, sedangkan pengkalibrasian dengan ML 2-P selalu menghasilkan SB yang berada di antara keduanya.

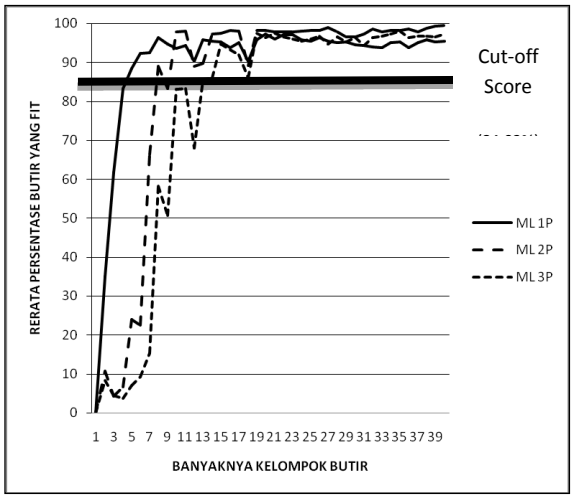
Nilai rerata pada Gambar 2 kemudian digunakan untuk melakukan estimasi interval terhadap rerata persentase butir yang *fit* pada populasi dengan *significance level* sebesar 5%. Batas bawah dari estimasi interval dimaksud kemudian dijadikan *cut-off score* yang posisinya dapat dilihat pada Gambar 4,

Gambar 5, dan Gambar 6. Gambar-gambar dimaksud juga memperlihatkan hubungan antara banyaknya kelompok butir dan rerata persentase butir yang *fit* pada setiap model yang digunakan untuk membangkitkan data yaitu ML 1-P, ML 2-P, dan ML 3-P secara berurutan.



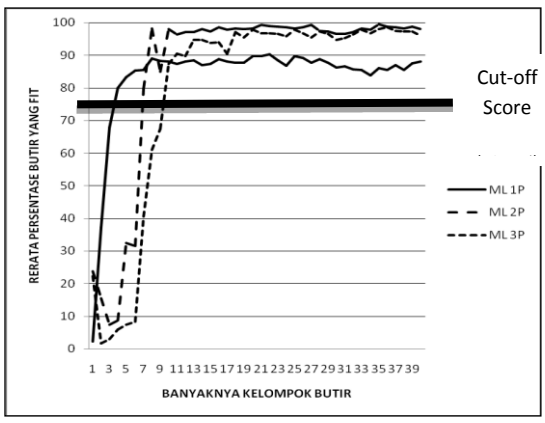
Gambar 3. Simpangan Baku Persentase Butir yang *Fit* untuk Data yang Dibangkitan dengan ML 1-P, ML 2-P, dan ML 3-P

Gambar 4 memperlihatkan bahwa ML 3-P mampu menghasilkan rerata persentase butir yang *fit* di atas skor batas maksimum jika banyaknya kelompok butir (faktor yang terekstraksi) sebanyak 13 atau lebih. Sementara itu ML 2-P mampu menghasilkan rerata persentase butir yang *fit* di atas skor batas maksimum jika banyaknya kelompok butir sebanyak 10 atau lebih. Untuk ML 1-P, rerata persentase butir yang *fit* di atas skor batas maksimum dapat tercapai jika banyaknya kelompok butir adalah 5 atau lebih. Karena semakin sedikit kelompok butir berarti rerata persentase pelanggaran Asumsi ILB semakin besar, maka hal ini berarti ML 1-P lebih *robust* dibandingkan dengan ML 2-P dan ML 3-P, jika data dibangkitkan dengan ML 1-P. Hasil ini sejalan dengan Gambar 2 dan Gambar 3.

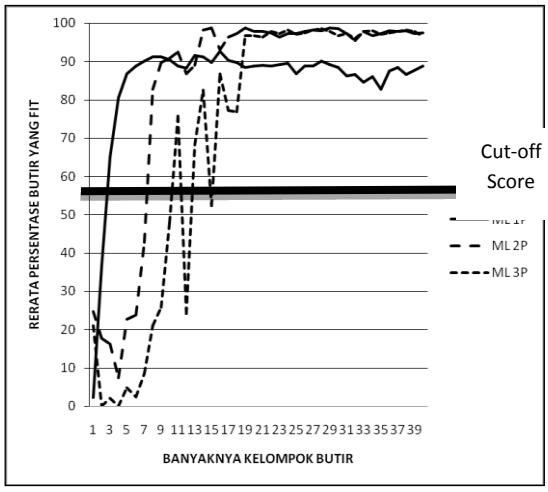


Gambar 4. Hubungan Banyaknya Kelompok Butir dan Rerata Persentase Butir yang *Fit* untuk Data yang dibangkitkan dengan ML 1-P

Gambar 5 memperlihatkan bahwa ML 3-P mampu menghasilkan rerata persentase butir yang *fit* di atas skor batas maksimum jika banyaknya kelompok butir sebanyak 10 atau lebih. Sementara itu ML 2-P mampu menghasilkan rerata persentase butir yang *fit* di atas skor batas maksimum jika banyaknya kelompok butir sebanyak 7 atau lebih. Untuk ML 1-P, rerata persentase butir yang *fit* di atas skor batas maksimum dapat tercapai jika banyaknya kelompok butir adalah 4 atau lebih. Hal ini berarti ML 1-P kembali terbukti lebih *robust* disbandingkan dengan ML 2-P dan ML 3-P, jika data dibangkitkan dengan ML 2-P. Hasil ini sejalan dengan Gambar 3.



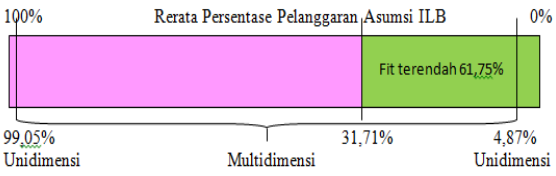
Gambar 5. Hubungan Banyaknya Kelompok Butir dan Rerata Persentase Butir yang *Fit* untuk Data yang dibangkitkan dengan ML 2-P



Gambar 6. Hubungan Banyaknya Kelompok Butir dan Rerata Persentase Butir yang *Fit* untuk Data yang dibangkitkan dengan ML 3-P

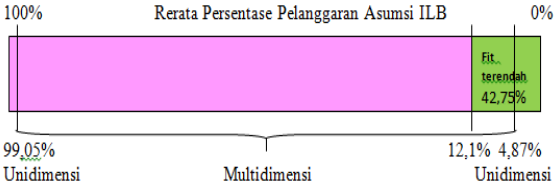
Gambar 6 memperlihatkan bahwa ML 3-P mampu menghasilkan rerata persentase butir yang *fit* di atas skor batas maksimum jika banyaknya kelompok butir (faktor yang terekstraksi) sebanyak 11 atau lebih. Sementara itu ML 2-P mampu menghasilkan rerata persentase butir yang *fit* di atas skor batas maksimum jika banyaknya kelompok butir sebanyak 8 atau lebih. Untuk ML 1-P, rerata persentase butir yang *fit* di atas skor batas maksimum dapat tercapai jika banyaknya kelompok butir adalah 3 atau lebih. Hal ini berarti ML 1-P tetap lebih *robust* dibandingkan dengan ML 2-P dan ML 3-P, jika data dibangkitkan dengan ML 1-P. Hasil ini juga sejalan dengan Gambar 2 dan Gambar 3.

Selanjutnya, berdasarkan hasil analisis faktor eksploratori dan kombinasi hasil dari rerata persentase butir yang *fit* untuk data yang dibangkitkan dengan ML 1-P, ML 2-P, dan ML 3-P didapatkan hal-hal sebagaimana tercantum pada Gambar 7, Gambar 8, dan Gambar 9 berikut ini.



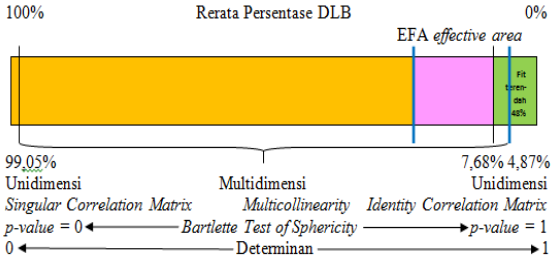
Gambar 7. Posisi *Robustness* ML 1-P

Gambar 7 memperlihatkan bahwa, jika tanpa memandang model yang digunakan untuk membangkitkan data, maka kalibrasi dengan ML 1-P mampu mentoleransi pelanggaran Asumsi ILB sebesar 31,71% dengan minimum butir yang *fit* rata-rata sebesar 61,75%.



Gambar 7. Posisi *Robustness* 2-PLM

Gambar 8 memperlihatkan bahwa, jika tanpa memandang model yang digunakan untuk membangkitkan data, maka kalibrasi dengan ML 2-P mampu mentoleransi pelanggaran Asumsi ILB sebesar 12,1% dengan minimum butir yang *fit* rata-rata sebesar 42,75%.



Gambar 8. Posisi *Robustness* ML 3-P

Gambar 9 memperlihatkan bahwa, jika tanpa memandang model yang digunakan untuk membangkitkan data, maka kalibrasi dengan ML 3-P mampu mentoleransi pelanggaran Asumsi ILB sebesar 7,68% dengan minimum butir yang *fit* rata-rata sebesar 48%. Dengan demikian, kembali terbukti bahwa ML 1-P lebih *robust* terhadap pelanggaran Asumsi ILB dibandingkan dengan ML 2-P maupun ML 3-P.

Selanjutnya, jika dibuat kategori-kategori skor bagi pelanggaran Asumsi ILB dengan berdasarkan dampaknya terhadap struktur dari matriks data, maka dapat dibuat tiga kategori sebagaimana yang tercantum pada Tabel 1. Kategori pertama adalah Pelanggaran Ringan, di mana pe-

langgaran Asumsi ILB tidak mengakibatkan terjadinya pelanggaran Asumsi Unidimensionalitas. Hal ini terjadi jika pelanggaran Asumsi ILB tidak lebih besar dari 4,87%. Kategori kedua adalah Pelanggaran Sedang, di mana pelanggaran Asumsi ILB hanya mengakibatkan terjadinya pelanggaran Asumsi Unidimensionalitas. Hal ini terjadi jika pelanggaran Asumsi ILB berada di antara 4,87% - 22,83%. Sementara itu kategori ketiga adalah Pelanggaran Berat, di mana pelanggaran Asumsi ILB tidak hanya mengakibatkan terjadinya pelanggaran Asumsi Unidimensionalitas tetapi juga membuat *data matrix is not positive definite*. Hal ini terjadi jika pelanggaran Asumsi ILB lebih besar atau sama dengan 22,83%.

Tabel 1. Kategori Skor Pelanggaran Asumsi ILB

No	Kategori Pelanggaran	Kriteria
1	Ringan (≤4,87%)	- Tidak mengakibatkan pelanggaran Asumsi Unidimensionalitas
2	Sedang (>4,87% - < 22,83%)	- Mengakibatkan pelanggaran Asumsi Unidimensionalitas
3	Berat (≥22,83%)	- Mengakibatkan pelanggaran Asumsi Unidimensionalitas. - Menghasilkan matriks yang "*not positive definite*".

Berdasarkan Tabel 1 dan Gambar 7, maka ML 1-P masih cukup *robust* terhadap pelanggaran Asumsi ILB yang termasuk dalam kategori berat. Sementara itu berdasarkan Tabel 1, Gambar 8, dan Gambar 9 dapat diketahui bahwa ML 2-P dan ML 3-P *robust* terhadap pelanggaran Asumsi ILB yang masih dalam kategori sedang. Dengan demikian, kembali dapat dibuktikan bahwa ML 1-P lebih *robust* terhadap pelanggaran Asumsi ILB dibandingkan dengan ML 2-P maupun ML 3-P.

Jika dilakukan analisis secara inferensia dengan menggunakan Model Linear yang Digeneralisasi (GLZ), maka didapatkan hasil sebagaimana yang tercantum pada Tabel 2 dan Tabel 3.

Tabel 2. Pengujian Efek Model dengan GLZ

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	112,373	1	,001
GEN	1,062	2	,588
CAL	14,834	2	,001
GEN * CAL	,709	4	,950

Tabel 2 memperlihatkan bahwa selain *intercept*, faktor yang berpengaruh terhadap rerata persentase butir yang *fit* adalah model yang digunakan untuk kalibrasi (CAL). Sementara itu model yang digunakan untuk membangkitkan data (GEN) terbukti tidak memiliki berpengaruh terhadap rerata persentase butir yang *fit*. Hal ini berarti ML 1-P *robust* terhadap pelanggaran Asumsi *Homogeneity of Discrimination* dan *Zero Lower Asymptote*, sedangkan ML 2-P *robust* terhadap pelanggaran Asumsi *Zero Lower Asymptote*.

Tabel 3. Estimasi Parameter dengan GLZ

Parameter	*Hypothesis Test*		
	Wald Chi-Square	df	Sig.
CAL 1	6,815	1	,009
CAL 2	1,693	1	,193
CAL 3*	.	.	.

** this parameter is redundant*

Tabel 3 memperlihatkan bahwa signifikansi terkecil yaitu 0,009 (jauh lebih kecil dari 0,05) terdapat pada kalibrasi dengan menggunakan ML 1-P, diikuti oleh ML 2-P. Sementara itu untuk penggunaan ML 3-P, nilai *Wald Chi-Square*, *df*, maupun signifikansinya tidak dapat dihasilkan karena para-

meter ini bersifat *redundant*. Dengan demikian, berdasarkan hasil analisis inferensia dengan menggunakan GLZ kembali dibuktikan bahwa ML 1-P lebih *robust* terhadap pelanggaran Asumsi ILB dibandingkan de-ngan ML 2-P maupun ML 3-P.

Simpulan dan Saran

Simpulan

Berdasarkan paparan pada temuan dan diskusi dapat disimpulkan beberapa hal berikut ini.

1. ML 1-P, ML 2-P, dan ML 3-P tidak sepenuhnya (100%) *robust* terhadap pelanggaran Asumsi ILB.
2. Berdasarkan batas bawah estimasi interval dengan tingkat signifikansi sebesar 5% sebagai skor batas secara umum dapat disimpulkan beberapa hal sebagaimana tercantum berikut ini: (a) ML 1-P mampu mentoleransi pelanggaran Asumsi ILB rata-rata sebesar 31,71% (kategori pelanggaran berat) dengan butir *fit* yang terendah adalah rata-rata sebesar 61,75%; (b) ML 2-P mampu mentoleransi pelanggaran Asumsi ILB rata-rata sebesar 12,1% (kategori pelanggaran sedang) dengan butir *fit* yang terendah adalah rata-rata sebesar 42,75%; (c) ML 3-P mampu mentoleransi pelanggaran Asumsi ILB rata-rata sebesar 7,68% (kategori pelanggaran sedang) dengan butir *fit* yang terendah adalah rata-rata sebesar 48%.
3. Simpulan dari analisis deskriptif di atas memperlihatkan bahwa model yang paling *robust* terhadap pelanggaran Asumsi ILB adalah ML 1-P, diikuti oleh ML 2-P, dan terakhir adalah ML 3-P. Hasil ini juga sama dengan hasil analisis inferensia menggunakan Model Linear yang digeneralisasi.
4. Hasil analisis inferensia lainnya memperlihatkan model yang digunakan untuk membangkitkan data tidak berpengaruh terhadap rerata persentase butir yang *fit*. Hal ini berarti bahwa ML 1-P *robust* terhadap pelanggaran Asumsi *homogeneity of discrimination* dan *pseudo chance level* = 0, sedangkan ML 2-P *robust* terhadap pelanggaran Asumsi *pseudo chance level* = 0, khususnya pada penelitian ini yang menggunakan butir sebanyak 40 dan simuli sebanyak 500.

Saran

Praktisi TRB hendaknya memperhatikan hal-hal berikut ini.

1. Memeriksa banyaknya faktor yang terekstraksi sebelum melakukan kalibrasi dengan menggunakan ML 1-P, ML 2-P, atau ML 3-P.
2. Memeriksa pelanggaran asumsi ILB terutama jika kalibrasi menghasilkan butir yang tidak *fit* melebihi 58,99%.
3. Secara umum dapat mengabaikan persoalan asumsi ILB jika kalibrasi menghasilkan: (1) butir yang *fit* ≥ 84,02% un-tuk ML 1-P, (2) butir yang *fit* ≥ 76,71% untuk ML 2-P, serta (3) butir yang *fit* ≥ 58,99% untuk ML 3-P.
4. Untuk bidang-bidang ilmu tertentu pemilihan model untuk mengkalibrasi data dapat menggunakan acuan umum berdasarkan hasil penelitian Scott & Ip (2002, p.2) mengenai data *The National Assessment of Educational Progress* (NAEP) sebagai berikut; (1) Untuk Matematika, dengan rerata pelanggaran Asumsi ILB sebesar 12%, maka lebih baik melakukan kalibrasi dengan ML 2-P dan ML 1-P; (2) Untuk Ilmu Pengetahuan Alam dengan rerata pelanggaran Asumsi ILB sebesar 74%, maka penggunaan ML 3-P, ML 2-P, dan ML 1-P sebaiknya tidak dilakukan karena hasil estimasi parameter-parameternya akan bias bahkan dapat memberikan hasil yang tidak valid. Untuk kasus seperti ini sebaiknya digunakan MRB Multidimensional atau Model-model Respons Teslet (*Testlet Response Models*). Namun karena model-model ini pada dasarnya berbasis kategorik maka lebih teliti jika menggunakan MRB Kondisional, MRB Efek Random, atau MRB Marjinal yang khusus untuk menangani pelanggaran Asumsi ILB yang berbasis numerik; (3) Untuk Kemampuan Memahami Bacaan

dengan rerata pelanggaran Asumsi ILB sebesar 100%, maka penggunaan ML 3-P, ML 2-P, ML 1-P, MRB Multidimensional, maupun Model Respons Testlet juga sebaiknya tidak dilakukan. Untuk kasus seperti ini MRB yang layak digunakan adalah MRB Kondisional, MRB Efek Random, atau MRB Marjinal. Karena pelanggaran Asumsi ILB sulit dihindari sepenuhnya, maka penggunaan tiga MRB yang dicantumkan belakangan layak untuk semua kondisi kecuali untuk kondisi dimana Asumsi ILB terpenuhi seutuhnya (pelanggarannya = 0%). Namun karena persentase DLB pada tiga bidang di atas berdasarkan data NAEP, maka acuan ini hanya merupakan acuan kasar saja. Sebaiknya, menggunakan hasil penelitian yang relevan dan akan lebih baik lagi jika dilakukan pemeriksaan pelanggaran Asumsi ILB sebelum melakukan kalibrasi dengan menggunakan MRB tertentu.

5. Tidak perlu melakukan pembatasan penggunaan ML 1-P, ML 2-P, maupun ML 3-P hanya untuk jenis instrumen tertentu (tes) saja.

Peneliti TRB dapat melakukan penelitian lanjutan dengan: (a) fokus pada besaran DLB yang lebih rinci terutama pada kelompok data sebanyak 4 atau kurang; (b) menggunakan distribusi parameter simuli dan parameter butir yang berbeda serta dengan interval nilai parameter yang berbeda; (c) menggunakan banyaknya simuli dan butir yang lebih bervariasi.

Daftar Pustaka

Ackerman, T. (September 1987). The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence. *ACT Research Report Series*, 87-14.

Anderson, T. W. (1954). On estimation of parameters in laten structure analysis. *Psychometrika*, 19 (1), 1 – 10.

Andrich, D. (2008). Relationships between the Thurstone and Rasch approach to item scaling. DalamS. Gorard (Ed.), *Quantitative research in education*: *Key techniques for education research* (pp. 66-78). London: 2008.

Antal, J. (2003). Fit indices for the rasch model. *Dissertation*, Ohio: The Ohio State University.

Balazs, K. & De Boeck, P. (2007). Detecting local item dependence stemming for minnor dimensions. *Interuniversity Attraction Pole* (IAP) *Statistics Network Technical Report Series*, 0684.

Bond, T. G. & Fox, C. H. (2007). *Applying the rasch model*: *Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum, Inc.

Bradlow, E. T. Wainer, H. & Wang, X. (January 1998). A bayesian random effects model for testlets. *Educational Testing Service* (ETS) *Research Report*, RR-98-3.

Braeken, J. & Tuerlinckx, F. (2009). Investigating latent constructs with item response models: A MATHLAB IRTm toolbox. *Behavior Research Methods*, 41, 1127-1137.

Braeken, J. Tuerlinckx, F. & De Boeck, P. (Juni 2005). A copula model for residual dependency in IRT models. *Interuniversity Attraction Pole* (IAP) *Statistics Network Technical Report Series*, 0534.

Brown, J. D. (2005). *Testing in language programs*: *A comprehensive guide to english language assessment*. New York: McGraw-Hill Companies, Inc.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

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

Edwards, M. C. (2009). An introduction to item response theory using the need cognition scale. *Social and Personality Psychology Compass*, 3, 507-529.

Embretson, S. E. (1996).The new rules of measurement. *Psychological Assessment*, 8, 341-349.

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

Fox, J. (2008). *Applied regression analysis and generalized linear models*. (2nd ed.) Thousand Oaks, CA: Sage Publications, Inc.

Fox, J. P. (2010). *Bayesian item response modeling*: *Theory and applications*. New York: Springer.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, 78, 350-365.

Freund, J. E. (2004). *Mathematical statistics with applications*. Upper Saddle River, NJ: Pearson Education International.

Goodman, J. T. & Luecht, R. M. (August 2009). An examination of the magnitude of residual covariance for complex performance assessments under various scoring and scaling methods. *American Institute of Certified Public Accountants* (*AICPA*) *Technical Report Series*, W0901.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *An NCME Instructional Module*, 253-262. Diakses tanggal 5 Juli 2010, dari: http://ncme.org/linkservid/ 6696808 0-1320-5CAE-6E4E546A2E4FA9E 1 /showMeta/0/

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

Hambleton, R. K. Swaminathan, H. & Rogers, H. J. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Hulin, C. L. Drasgow, F. & Parsons, C. K. (1983). *Item response theory*: *Application to psychological measurement*. Homewood, IL: The Dorsey Professional Series.

Huynh, H. Michels, H. R. & Ferrara, S. (April 1995). A comparison of three statistical procedures to identify clusters of items with local dependency. *Paper presented at the Annual Meeting of National Council on Measurement in Education*, di San Francisco.

Ip, E. H. Wang, Y. J. De Boeck, P. et al. (2004). Locally dependent latent trait model for polytomous responses with application to inventory of hostility. *Psychometrika*, 69, 191-216.

Jiao, H. & Kamata, A. (April 2003). Model comparisons in the presence of local item dependence. *Paper presented at the Annual Meeting of The American Educational Research Association* (AERA), di Chicago.

Jiao, H. Kamata, A. Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, 49, 82-100.

Johnson, M. S. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software*, 20, 1-24.

Kim, D. De Ayala, R. J. Ferdous, A. A. et al. (2007). *Assessing relative performance of local item dependence (LID) indexes*. Diakses tanggal 5 Juli 2010, dari http:// www.Measuredprogress.org/ resources/psychometrics/framework /materials/07/AERA.NCME/Assess ingRelativePerformnace.pdf.

Lord, F. M. (1952).*A theory of test scores*. New York: Educational Testing Service.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Adison-Wesley.

Mislevy, J. L. (2011). Detecting local item dependence in polytomous adaptive data. *Dissertation*. Maryland: University of Maryland.

Mokken, R. J. (1997). Nonparametric models for dichotomous response. Dalam W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York: Springer-Verlag.

Orlando, M. (2008). *Critical issues to address when applying item response theory* (*IRT*). Diakses tanggal 5 Juli 2010, dari

http://outcomes.cancer.gov/conference/irt/orlando.pdf.

Panther, A. T. & Reeve, B. B. (2002). Assessing tobacco beliefs among youth using item response theory models. *Drug and Alcohol Dependence*, 68, S21-S39.

Pommerich, M. & Ito, K. (March 2008). An examination of the properties of local dependence measures when applied to adaptive data. Paper presented at *the Annual Meeting of the National Council on Measurement in Education* (NCME), di New York.

Ramsay, J. O. (1997). A functional approach to modeling test data. Dalam W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 381-394). New York: Springer-Verlag.

Reese, L. M. (April 1995). The impact of local dependencies on some LSAT outcomes. *Law School Admission Council* (LSAC) *Statistical Report*, 95-02.

Scott, S. L. & Ip E. H. (2002). Empirical bayes and item-clustering effects in a latent variable hierarchical model: A case study from the national assessment of educational progress. *Journal of the American Statistical Association*, 97, 1-11.

Sijtsma, K. & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika*, Vol. 33, No. 1, 75-102.

Spray, J. A. (1997). Multiple-attempt, single-item response models. Dalam W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York: Springer-Verlag.

Stark, S. Chernyshenko, O. S. & Drasgow, F. (April 2002). Investigating the effects of local dependence on the accuracy of IRT ability estimation. *American Institute of Certified Public Accountants* (AICPA) *Technical Report Series*, 15.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: John Wiley & Sons, Inc.

Traub, R. E. (1983). Apriori considerations in choosing an item response model. Dalam R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 1-23). Vancouver, BC: Educational Research Institute of British Columbia.

Tuerlinckx, F. & De Boeck, P. (2001). Non-modeled item interactions lead to distorted discrimination parameters: A case study. *Methods of Psychological Research Online*, 6, 159-174.

van der Linden, W. J. & Hambleton, R. K. (1997) Nonparametric models. Dalam W. J. van der Linden, & R. K. Hambleton, (Eds.). *Handbook of modern item response theory* (pp. 347-349). New York: Springer-Verlag.

Wainer, H. Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wang, W. C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, 29, 296-318.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago, IL: Mesa Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8 (2), 125-145

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2003). Effects of local item dependence on the validity of IRT item, test, and ability statistics. *Medical Council Admission Test* (MCAT) *Report Series*.

Zwinderman, A. A. (1997). Response models with manifest predictors. Dalam W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. s351-367). New York: Springer-Verlag.