PENYETARAAN VERTIKAL MODEL KREDIT PARSIAL SOAL MATEMATIKA SMP

Sugeng
Universitas Mulawarman
kenduk_s@yahoo.com

Abstrak

Penelitian ini bertujuan menemukan ukuran sampel minimum, pengaruh panjang tes, panjang tes *anchor* minimum, dan metode penyetaraan tes dalam penyetaraan vertikal model kredit parsial soal Matematika SMP menggunakan *common-item nonequivalent groups design*. Pembangkitan data melibatkan variasi peringkat kelas terhadap ukuran sampel (300; 600; 1000), panjang tes (10; 20), dan distribusi kemampuan ($N(0,1)$, $N(1,1)$) sebanyak 50 replikasi menggunakan Program *WinGen*2. Penyetaraan vertikal melibatkan (a) panjang tes *anchor* 2, 3, 4, 5, dan 8 butir (panjang tes 20 butir); dan (b) panjang tes *anchor* 2, 3, 4, dan 5 butir (panjang tes 10 butir). Kriteria pengujian keakuratan penyetaraan menggunakan *RMSD* dan *RMSE*. Hasil penelitian menunjukkan: (1) Penyetaraan vertikal pada sampel 300 memiliki rata-rata *RMSD* dan *RMSE* cukup kecil untuk semua situasi; (2) Keakuratan penyetaraan meningkat seiring meningkatnya panjang tes; (3) Dengan rentang panjang tes *anchor* 25% sampai 30% untuk butir politomus, penyetaraan vertikal model kredit parsial memerlukan panjang tes *anchor* minimum 5 untuk panjang tes 20 butir dan 3 untuk panjang tes 10 butir; dan (4) Metode *Mean/Mean* cenderung lebih akurat, dalam penyetaraan vertikal *IRT* butir tes Matematika model kredit parsial diikuti *Stocking-Lord*, *Mean/Sigma*, dan *Haebara*.

Kata kunci: *penyetaraan vertikal, model kredit parsial, tes anchor, kalibrasi, RMSD*, *RMSE*

VERTICAL EQUATING USING PARTIAL CREDIT MODEL FOR JUNIOR HIGH SCHOOL MATHEMATICS TESTS

Sugeng
Mulawarman University
kenduk_s@yahoo.com

Abstract

This study aims to find the minimum sample size, the effect of the test length, the minimum length of the anchor test, and an accurate test equating method of vertical equating using the partial credit model of Mathematisc for Junior High School (JHS). This study used the common-item nonequivalent groups design. The data were generated using the WinGen2 Program involving a grades variation with the test length factor (20 and 10), the sample size factor (300, 600, and 1000), and the ability distribution factor (N(0,1), N(1,1)). Vertical equating involving (a) the anchor test lengths of 2, 3, 4, 5, and 8 items for the test length of 20 items; and (b) the anchor test lengths of 2, 3, 4, and 5 items for the test length of 20 items. The test equating accuracy employed RMSD and RMSE. The results are: (1) minimum sample size is 300 has relatively small means of RMSD and RMSE in all situations. (2) The equating accuracy increases as the test length increases. (3) At an anchor test length ranging from 25 % to 30 % for polytomous items, the partial credit model needs an anchor test with a minimum length of 5 for a test length of 20 items and 3 for a test length of 10 items. (4) The Mean/Mean method tends to be more accurate followed by the Stocking-Lord (S-L), the Mean/Sigma, and the Haebara method respectively.

Key word: *vertical equating, the partial credit model, anchor test, calibration, RMSD, RMSE*

Pendahuluan

Pemerintah melalui Badan Standar Nasional Pendidikan (BSNP) berusaha meningkatkan kualitas pendidikan dengan mengembangkan standar nasional pendi-dikan. Untuk itu diperlukan suatu program pengukuran berskala nasional/daerah yang tepat. Pelaksanaan program berskala nasional melibatkan person banyak, koreksi cepat, dan objektif. Kelemahannya, jika terjadi kebocoran pada salah satu wilayah Indonesia, misal Indonesia bagian timur, dapat mengakibatkan gagalnya tujuan pelaksanaan program pengukuran. Untuk mengantisipasi terjadinya kebocoran dan untuk keamanan pelaksanaan pengukuran, misal program Ujian Nasional, perlu dilakukan penyetaraan.

Pengembangan sekaligus perbaikan pembelajaran bidang studi Matematika secara internal sekolah sangat mendukung pencapaian standar nasional bidang Matematika. Untuk itu, diperlukan adanya penelusuran tingkat kemampuan tiap jenjang kelas suatu jenjang pendidikan, misal SMP, melalui penyetaraan vertikal.

Model *IRT* politomus *Partial Credit Model* (*PCM*) diaplikasikan untuk butir soal Matematika. Pelaksanaan penyetaraan vertikal skor Matematika *PCM* melibatkan ukuran sampel, panjang tes, panjang tes *anchor*, desain, dan empat metode (*Mean/ Mean*, *Mean/Sigma*, *Haebara*, dan *Stocking-Lord*). Dari keempat metode tersebut dipilih satu metode yang menunjukkan hasil penyetaraan paling akurat.

Teori Respons Butir (*Item Response Theory; IRT*) merupakan suatu pendekatan pengukuran yang berdasar respons-respons butir soal untuk mengetahui karakteristik laten suatu objek (Hulin, Drasgow, & Parsons, 1983:15). Kinerja peserta tes dalam merespons suatu butir soal tes bergantung kepada kemampuan yang dimilikinya; semakin tinggi tingkat kemampuan yang dimiliki akan semakin baik kinerja yang ditampilkan peserta tes sebagaimana digambarkan dengan kurva yang monoton naik.

Model kredit parsial (*Partial Credit Model, PCM)* adalah salah satu model *IRT* politomus, dikembangkan oleh Masters (1982) berdasarkan model Rasch (model 1-PL) yang respons butirnya dikotomus menjadi model yang responnya lebih dari dua kategori terurut (politomus). Model

IRT politomus *PCM* mengasumsikan bahwa semua butir soal memiliki indeks diskriminasi sama (Embretson & Reise, 2000:106). Respons terhadap suatu butir soal *j* diklasifikasikan ke dalam (m_j + 1) kategori terurut. Skor kategori pada butir *j* adalah bulat dan berurutan, dinyatakan *x* dan harga *x* adalah 0, 1, 2, …, m_j. Suatu skor kategori merepresentasikan banyaknya langkah yang sukses diselesaikan.

Probabilitas respons individu *i* dengan tingkat kemampuan θ memperoleh skor kategori *k* pada suatu butir soal *j*, dinyatakan dalam model *PCM* (Masters, 1982):

$$P_{ijk} = P_{jk}(\theta_i) = \frac{\exp \sum_{k=0}^{x} (\theta_i - b_{jk})}{\sum_{h=0}^{m_j} \exp\left[\sum_{k=0}^{h} (\theta_i - b_{jk})\right]} \quad x = 0, 1, 2, 3, \ldots, m_j \qquad (1)$$

dengan $\sum_{k=0}^{0} (\theta_i - b_{jk}) = 0$; b_{jk} menyatakan parameter tingkat kesulitan butir *j* berkenaan dengan skor kategori *k*, dan *k* sebagai skor kategori tertinggi yang tercapai.

Penyetaraan merupakan proses statistis yang digunakan untuk menyesuaikan skor pada instrumen-instrumen tes sedemikian sehingga skor-skor itu dapat digunakan saling bertukar (Kolen & Brennan, 1995: 2, 2004: 2), atau skor-skor instrumen tes yang satu dapat diberlakukan pada instrumen tes lainnya (Peterson, Kolen, & Hoover, 1989). Penyetaraan vertikal sebagai penyetaraan yang melibatkan dua atau lebih instrumen tes yang mengukur *trait* sama, dengan tingkat kesulitan soal berbeda, distribusi kemampuan peserta berbeda, kelompok peserta tes berasal dari level kelas berbeda (Hambleton & Swaminathan, 1985: 197; Crocker & Algina, 1986: 473). Proses penyetaraan vertikal *IRT* memerlukan desain, *common scale*, dan skor tes (Cook & Eignor, 1991).

Penyetaraan *IRT* menggunakan parameter butir hasil estimasi. Proses penentuan parameter-parameter dari fungsi respons suatu butir yang memuat parameter butir dan parameter person melalui kalibrasi (*Standards for Educational and Psychological Testing*, 1999: 172). Hasil penyetaraan bergantung kepada metode estimasi parameter yang digunakan (Ogasawara,

2001). Program *QUEST* menggunakan *the unconditional maximum likelihood procedure* (*UCON*), yakni sebagai salah satu prosedur untuk mengestimasi parameter butir *PCM* (Wright & Masters, 1982: 86).

Soal Matematika non-rutin sesuai diterapkan untuk memaksimalkan kemampuan berpikir dan pemecahan masalah topik Matematika yang isinya melibatkan permasalahan kehidupan nyata (*NTCM,* 2000 dalam Kennedy, Tipps, & Johnson, 2008: 3). Banyaknya langkah penyelesaian soal Matematika oleh siswa tidak dapat digunakan sebagai pedoman dalam menentukan skor suatu butir. Akibatnya, estimasi terhadap parameter butir tertentu tidak berharga tunggal sehingga butir soal tidak dapat dianalisis lebih lanjut. Oleh karenanya, perlu dilakukan pengelompokan terhadap variasi komponen sejenis (Ferrara & Walker-Bartnick, 1989) dalam penskalaan *PCM*.

Pemilihan butir soal dilakukan dengan mengamati fungsi informasi (Dodd & Ayala, 1994) berdasarkan kurva respons butir. Secara grafis, tingkat kesulitan tiap langkah teramati pada θ tertentu dalam interval $-3 \leq \theta \leq +3$. Pada kurva respons butir, semakin *ability* meningkat ke arah level moderat, probabilitas jawaban salah semakin menurun, dan jawaban benar secara parsial meningkat (Yen & Fitzpatrick, 2006: 115).

Penyetaraan vertikal soal Matematika model *IRT* politomus *PCM* menggunakan metode *Mean/Mean*, *Mean/Sigma*, *Haebara*, dan *Stocking-Lord*. Koefisien penyetaraan (α, β) pada metode *Mean-Mean* ditentukan melalui rata-rata hasil estimasi parameter tingkat kesulitan dan diskriminasi dari *n* butir pada tes *anchor*. Metode *Mean/Sigma* menggunakan rata-rata dan standar deviasi dari estimasi parameter tingkat kesulitan butir pada *n* butir *common-item*, dihitung menurut *m* kategori skor setiap kelompok.

Haebara (1980) mengembangkan metode kuadrat terkecil untuk mentransformasi skala logistik. Koefisien penyetaraan metode *Haebara* diperoleh dengan meminimumkan rata-rata jumlah kuadrat dari selisih antara fungsi-fungsi respons butir yang dihasilkan butir-butir tes *anchor* dari dua tes. Fungsi kriteria pada metode *Haebara*:

$$F = \frac{1}{N}\sum_{i=1}^{N}\left\{\sum_{j=1}^{n}(T_{ij} - T^{*}{}_{ij})\right\}^{2} \qquad (2)$$

N adalah jumlah peserta *i* (*i*=1, 2, …, *N*), T_{ij} adalah skor sejati butir *j* (*j*=1, 2, …, *n*) tes *anchor* peserta *i*, dan T^*_{ij} adalah hasil transformasi skor sejati T_{ij} pada butir tes *anchor* ke-2 onto skala tes *anchor* ke-1. (Baker, 1993) mengembangkan fungsi kriteria nonlinear *F* untuk model *IRT* politomus berdasarkan fungsi kriteria *Haebara*.

Stocking-Lord (1983) memodifikasi metode *Haebara*. Koefisien penyetaraan (α dan β) pada metode *Stocking-Lord* diperoleh dengan cara meminimumkan rata-rata kuadrat jumlah dari selisih antara estimasi skor sejati fungsi-fungsi respons butir tes *anchor* dua tes. Menurut metode *Stocking-Lord*, fungsi kriteria *F* adalah

$$F = \frac{1}{N}\sum_{i=1}^{N}(T_i - T^*_i)^2 \qquad (3)$$

N adalah jumlah peserta *i* (*i*=1, 2, …, *N*); T_i adalah skor sejati, yakni jumlah dari probabilitas respons benar peserta *i* terhadap butir *j* tes *anchor*; dan T_i^* adalah hasil transformasi skor sejati T_i pada butir tes *anchor* ke-2 onto skala tes *anchor* ke-1. (Baker, 1993) mengembangkan fungsi kriteria nonlinear *F* model *IRT* politomus berdasarkan fungsi kriteria *Stocking-Lord*.

Terkait dengan hal di atas, penelitian ini dilaksanakan dengan tujuan menemukan ukuran sampel minimum, pengaruh panjang tes, panjang tes *anchor* minimum, dan metode penyetaraan tes dalam penyetaraan vertikal model kredit parsial. Tes yang digunakan adalah tes pada mata pelajaran matematika SMP.

Metode Penelitian

Penyetaraan vertikal melibatkan dua atau lebih instrumen tes yang berbeda tingkat kesulitan butirnya dan kelompok peserta berbeda peringkat kelasnya. Instrumen tes yang disetarakan mengukur *content* sama. Berarti, kedua peringkat kelas mempelajari bidang studi sama, butir-butir soal mengukur *content* sama, dan kelas lebih rendah menyelesaikan soal dengan materi kelas di atasnya, atau sebaliknya. Oleh karena itu, penyetaraan

vertikal diaplikasikan pada bidang studi yang memiliki karakteristik vertikal penalaran, bukan hafalan.

Penelitian ini menggunakan ukuran sampel bervariasi, yaitu 300 (kecil), 600 (medium), dan 1000 (besar). Pembangkitan data dengan Program *WinGen*2 (Han & Hambleton, 2007), melibatkan ukuran sampel dan peringkat kelas (VII, VIII, IX) yang dikondisikan menurut VII-300/VIII-300/IX-300; VII-600/VIII-600/IX-600; dan VII-1000/VIII-1000/IX-1000. Setiap kondisi dilakukan variasi menurut panjang tes 20 atau 10 butir dengan masing-masing kemampuan θ pada distribusi normal *N*(0,1) dan *N*(1,1) sebanyak 50 replikasi. Data simulasi kelas VIII (kelompok *reference*), dibangkitkan pada kondisi *N*(0,1); untuk kelas VII dan kelas IX dibangkitkan pada kondisi *N*(0,1) dan *N*(1,1). Penyetaraan vertikal *IRT PCM* melibatkan panjang tes *anchor*, dengan variasi 2, 3, 4, 5, 8 (panjang tes 20 butir); dan 2, 3, 4, 5 (panjang tes 10 butir). Proses penyetaraan vertikal memakai *common-item nonequivalent groups design* dan penentuan koefisien penyetaraan dengan Program *STUIRT* (Kim & Kolen, 2004).

Keakuratan metode penyetaraan diukur dengan (a) *Root Mean Square Difference* (*RMSD*) antara parameter hasil estimasi dan parameter hasil bangkitan; dan (b) *Root Mean Square Error* (*RMSE*) antara parameter butir hasil estimasi dan parameter sejatinya. Menurut Gifford & Swaminathan (1990), pengujian kualitas penyetaraan dapat dilakukan dengan menggunakan *RMSD* yang diperoleh dari *MSD*, aturannya:

$$\sum_{k=1}^{n}\frac{(m_k-\tau)^2}{r}=(m.-\tau)^2+\sum_{k=1}^{r}\frac{(m_k-m.)^2}{r} \qquad (4)$$

Mean Squared Difference (*MSD)* sebagai rata-rata kuadrat dari selisih antara estimasi τ pada replikasi ke *k* dan τ; m_k sebagai estimasi τ pada replikasi *k*; *m*. sebagai rata-rata estimasi τ pada replikasi *r*, dan τ sebagai parameter sejati. *MSD* dinyatakan sebagai jumlah dari *Bias* dan *Variance* parameter butir yang diestimasi. *RMSE* untuk parameter tingkat kesulitan *b*, (Kirisci, Hsu, & Yu, 2001); dengan *n* adalah banyaknya replikasi,

$$RMSE_b = \sqrt{\frac{\sum_{j=1}^{n} (\hat{b}_{ij} - \beta_i)^2}{n}} \quad (5)$$

$\hat{b}_{ij}$ adalah estimasi dari parameter tingkat kesulitan butir *i* pada replikasi *j*, β_i adalah parameter tingkat kesulitan sejati. Semakin kecil harga *RMSE* dan *RMSD* menunjukkan metode penyetaraan semakin akurat dan kualitas penyetaraan semakin baik.

Analisis Data Empiris

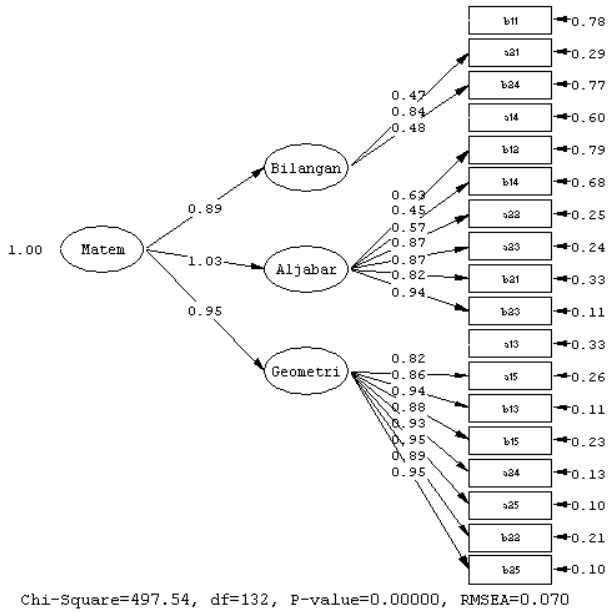
Dua butir soal Matematika model *IRT* politomus *PCM* (a_{11} dan a_{12}), secara grafis tidak memenuhi syarat sebagai soal *IRT* dan dikeluarkan dari instrumen tes. Jumlah butir soal kelas VII dan kelas IX masing-masing 18 butir, dan kelas VIII tetap 20 butir. Susunan kelompok butir soal setiap instrumen tes dengan teknik *overlapping* (Loyd & Hoover, 1979). Penyelidikan validitas konstruk dan unidimensi instrumen tes melibatkan keseluruhan data menurut masing-masing peringkat kelas. Hasil analisis *person-fit* terhadap keseluruhan respons, diperoleh kelas VII (560 dari 749 siswa), kelas VIII (577 dari 840 siswa), dan kelas IX (509 dari 749 siswa). Pemilihan person/respons dilakukan lebih dari satu kali dengan Program *QUEST* (Adams & Khoo, 1996).

Instrumen tes kelas VII berjumlah 18 butir. Penyelidikan validitas konstruk dan unidimensi menggunakan *Structural Equation Modeling* (*SEM*) (Hoe, 2008), ukuran sampel 560, *MA=PM*, *ME=WLS* (Jöreskog & Sörbom, 1996), dengan Program *LISREL* 8.51 dan perlu ukuran sampel minimum ½(18)(18-1)=153 (Gambar 1). Instrumen tes kelas VIII berjumlah 20 butir. Penyelidikan validitas konstruk dan unidimensi melibat-kan ukuran sampel 577; perlu ukuran sampel minimum ½(20)(20-1)=190 (Gambar 2). Instrumen tes kelas IX berjumlah 18 butir. Penyelidikan validitas konstruk dan unidi-mensi melibatkan sampel 509. Ukuran sampel minimum ½(18)(18-1)=153 (Gambar 3).

Hasil analisis *LISREL* kelas VII menunjukkan indeks kesesuaian *GFI* = 0,97; *RMSEA* = 0,070; *RMR* = 0,20 (*absolute*); *AGFI* = 0,96; *CFI* =

0.95 (*incremental*); *PGFI* = 0,75; rasio χ^2/d.f. = 3,77 (*parsimony*). Hasil analisis pada kelas VIII menunjukkan indeks kesesuaian *GFI* = 0,97; *RMSEA* = 0,079; *RMR* = 0,23 (*absolute*); *AGFI* = 0,96; *CFI* = 0.95 (*incremental*); dan *PGFI* = 0,77; rasio χ^2/d.f. = 4,62 (*parsimony)*. Hasil analisis untuk kelas IX menunjukkan indeks kesesuaian *GFI*= 0,97; *RMSEA* = 0,061; *RMR* = 0,17 (*absolute*); *AGFI* = 0,97; *CFI* = 0.95 (*incremental*); dan *PGFI* = 0,76; rasio χ^2/d.f. = 2,88 (*parsimony*). Kondisi tersebut membuktikan hasil pengujian model hipotetik konseptual instrumen tes Matematika kelas VII, kelas VIII, dan kelas IX didukung oleh data empiris. Dengan demikian, keseluruhan model pengukuran (*goodness of fit*) pada instrumen tes Matematika semua peringkat kelas dapat diterima.

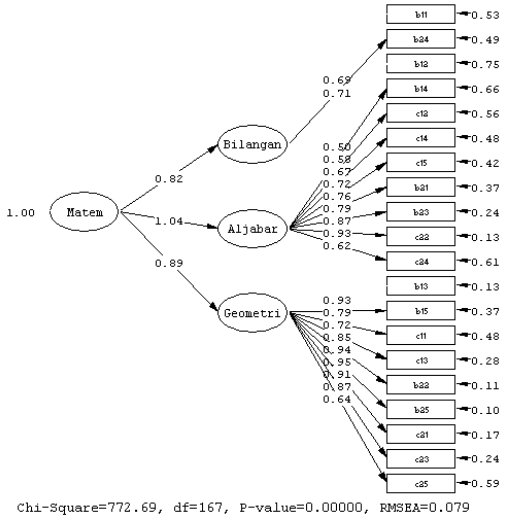
Menurut Gambar 1, terbukti bahwa variabel manifes (b_{11}, a_{21}, b_{24}) bagian dari variabel laten Bilangan; variabel manifes (a_{14}, b_{12}, b_{14}, a_{22}, a_{23}, b_{21}, b_{23}) bagian dari Aljabar; dan variabel manifes (a_{13}, a_{15}, b_{13}, b_{15}, a_{24}, a_{25}, b_{22}, b_{25}) bagian dari Geometri.



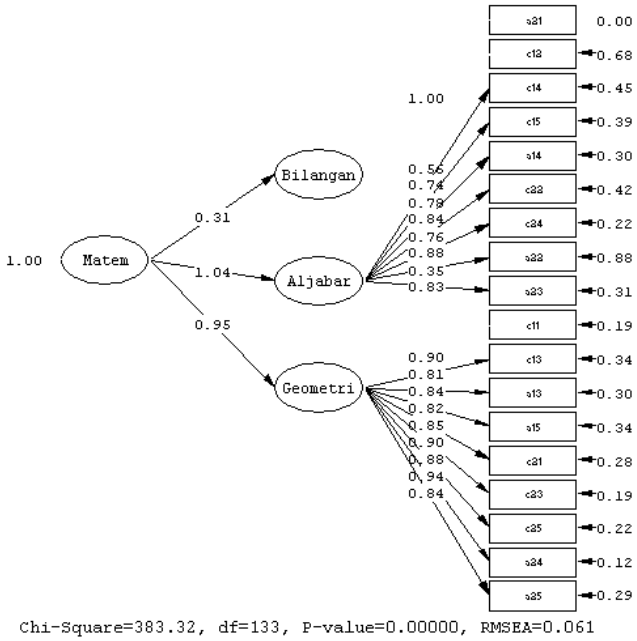
Gambar 1. Model Pengukuran Unidimensi Tes Matematika Kelas VII

Dengan demikian, secara teoretis variabel laten Bilangan, Aljabar, dan Geometri pada instrument kelas VII, masing-masing layak diukur dengan variabel-variabel manifesnya. Demikian juga untuk instrumen kelas VIII (Gambar 2) dan kelas IX (Gambar 3), masing-masing layak diukur dengan variabel-variabel manifesnya.

Hubungan antara variabel-variabel laten (Bilangan, Aljabar, dan Geometri) dan variabel laten Matem ditunjukkan dengan koefisien estimasi parameter *Gamma* (γ). Masing-masing analisis model pengukuran menunjukkan adanya bukti bahwa variabel laten Bilangan, Aljabar dan Geometri adalah bagian dari variabel Matem. Berarti, secara teoretis instrumen tes Matem pada kelas VII (Gambar 1) dapat diukur dengan variabel laten Bilangan, Aljabar, dan Geometri. Demikian juga untuk instrumen kelas VIII (Gambar 2) dan kelas IX (Gambar 3), secara teoretis instrumen tes Matem dapat diukur dengan variabel laten Bilangan, Aljabar, dan Geometri. Hasil ini memberikan indikasi bahwa validitas konstruk instrumen tes Matematika kelas VII, VIII, dan IX, masing-masing terbukti dan asumsi unidimensi terpenuhi. Dengan demikian, keseluruhan instrumen tes Matematika memenuhi kriteria validitas konstruk dan asumsi unidimensi.



Gambar 2. Model Pengukuran Unidimensi Tes Matematika Kelas VIII



Gambar 3. Model Pengukuran Unidimensi Tes Matematika Kelas IX

Analisis Data Simulasi

Analisis data simulasi melibatkan dua kelompok *Target* dan satu kelompok *Base* masing-masing berdistribusi kemampuan normal sama, yaitu *N*(0,1). Pada tahap berikut-nya, penyetaraan vertikal melibatkan dua kelompok *Target* dan satu kelompok *Base*, masing-masing memiliki distribusi kemampuan berbeda. Kelompok *Target* berdistribusi kemampuan *N*(0,1) dan kelompok *Base* berdistribusi kemampuan *N*(1,1).

Hasil Penelitian dan Pembahasan

Koefisien Penyetaraan

Derajat akurasi koefisien a (*slope*) lebih rendah daripada koefisien β (*intercept*). Hal ini ditunjukkan harga *RMSD* koefisien a lebih tinggi daripada

RMSD koefisien β untuk semua metode. Hasil perhitungan koefisien a untuk ketiga metode (*S-L*, *M/S*, *HA*) mendekati 1, bahkan koefisien a pada metode *M/M*, senantiasa 1 untuk semua panjang tes (10, 20) dengan berbagai ukuran sampel (300, 600, 1000) dan panjang tes *anchor*. Harga koefisien β untuk semua metode dan kondisi adalah mendekati 0. Hasil hitung koefisien a dan β ini mendekati harga teoretis koefisien penyetaraan yang diharapkan.

Ukuran Sampel

Ukuran sampel 1000 cenderung memiliki rata-rata *RMSD* dan *RMSE* paling kecil daripada ukuran sampel 600 ataupun 300 pada semua kondisi (pada N_i=10, 3x1x4x4x2x2x2=384 kondisi; N_i=20, 480 kondisi). Semakin bertambah besar ukuran sampel, rata-rata *RMSD* dan *RMSE* semakin kecil. Berarti, ukuran sampel berpengaruh terhadap pencapaian rata-rata *RMSD* ataupun *RMSE*. Semakin bertambah besar ukuran sampel, koefisien penyetaraan yang dihasilkan semakin akurat. Dengan demikian, ukuran sampel mempengaruhi kualitas penyetaraan vertikal model kredit parsial.

Pada ukuran sampel 300, rata-rata koefisien a mencapai harga mendekati 1 dan koefisien β mendekati 0. Pencapaian rata-rata *RMSD* koefisien a dan β masing-masing cukup kecil (0,095 dan 0,083 untuk N_i=20; 0,100 dan 0,075 untuk N_i=10). Selain itu, pencapaian rata-rata *RMSE* parameter tingkat kesulitan *b* dan b_i juga cukup kecil (0,022 dan 0,037 untuk *Ni*=20; 0,020 dan 0,047 untuk N_i=10). Berarti, ukuran sampel 300 sebagai ukuran sampel minimum memiliki rata-rata *RMSD* dan *RMSE* cukup kecil sehingga dapat menghasilkan kualitas penyetaraan yang baik.

Hasil ini mendukung bahwa peningkatan ukuran sampel membawa kepada hasil penyetaraan yang lebih akurat (Harris (1991); dan bahwa faktor ukuran sampel secara nyata berpengaruh terhadap keakuratan hasil penyetaraan (Nonny Swediati, 1997).

Panjang Tes

Rata-rata *RMSD* dan *RMSE* untuk panjang tes 20 butir cenderung lebih kecil daripada untuk panjang tes 10 butir pada semua kondisi. Berarti,

hasil penyetaraan menggunakan panjang tes 20 butir lebih akurat daripada dengan panjang tes 10 butir untuk semua metode. Panjang tes berpengaruh terhadap pencapaian *RMSD* dan *RMSE*. Panjang tes berpengaruh terhadap kualitas penyetaraan vertikal model kredit parsial.

Panjang Tes *Anchor*

Rata-rata *RMSD* dan *RMSE* semakin menurun seiring dengan penggunaan panjang tes *anchor* yang semakin meningkat. Berarti, panjang tes *anchor* berpengaruh terhadap pencapaian rata-rata harga *RMSD* dan *RMSE*. Dengan demikian, panjang tes *anchor* berpengaruh terhadap kualitas penyetaraan vertikal model kredit parsial. Panjang tes *anchor* pada ukuran sampel 1000 menghasilkan kualitas penyetaraan lebih baik daripada penyetaraan dengan ukuran sampel 600 atau 300.

The rule of thumb butir tes *anchor* model dikotomus antara 20%–25% (Hambleton, Swaminathan, & Rogers (1991: 135). sulit diterapkan pada model politomus, karena jumlah butirnya cenderung kecil. Hasil analisis varians panjang tes *anchor* menunjukkan tidak ada perbedaan antarpanjang tes *anchor* (p>0,05). Dengan mengambil rentang panjang tes *anchor* 25% – 30% untuk butir politomus, model kredit parsial dengan panjang tes 20 butir menggunakan panjang tes *anchor* minimum 5 butir dan panjang tes 10 butir menggunakan panjang tes *anchor* minimum 3 butir.

Metode Penyetaraan

Menurut hasil penelitian, 75% atau 18 dari 24 kelompok proses penyetaraan, menunjukkan keakuratan hasil penyetaraan cenderung ditentukan oleh metode *M/M*, *S-L*, *M/S*, dan *HA* secara berurutan. Secara keseluruhan, metode *M/M* cenderung mampu menghasilkan *RMSD* dan *RMSE* lebih kecil daripada metode *S-L*, *M/S* dan *HA*. Dengan demikian, untuk penyetaraan vertikal *IRT* butir tes Matematika politomus *PCM*, metode *M/M* lebih akurat daripada ketiga metode yang lain. Kondisi ini mendukung investigasi Baker & Karni (1991), bahwa metode *M/M* lebih bagus karena keadaan *mean* lebih stabil daripada standar deviasi, dan metode *M/M* hanya melibatkan unsur *mean*.

Penyetaraan vertikal skor butir *IRT* politomus *PCM* menghasilkan (a) koefisien β terdistribusi pada $0 \leq \beta \leq 1$ untuk semua metode, (b) koefisien a terdisribusi pada $0 \leq a \leq 1$ untuk metode *S-L*, *M/S*, *HA*; dan (c) koefisien a untuk metode *M/M* selalu 1, pada semua kondisi. Oleh karena itu, harga *RMSD* koefisien a untuk metode *M/M* senantiasa nol dan untuk ketiga metode lainnya bervariasi dari 0 sampai 1. Keadaan ini menyebabkan metode *M/M* sebagai metode yang lebih akurat daripada ketiga metode lainnya.

Distribusi Kemampuan Kelompok

Rata-rata *RMSD* dan *RMSE* pada distribusi kemampuan *N*(0,1) lebih kecil daripada *N*(1,1) untuk semua ukuran sampel (300, 600, 1000). Pada distribusi kemampuan kelompok *Target N*(0,1), rata-rata *RMSD* dan *RMSE* lebih kecil daripada kelompok *Target N*(1,1) dengan kelompok *Base* berdistribusi *N*(0,1). Berarti, distribusi kemampuan kelompok berpengaruh terhadap pencapaian rata-rata *RMSD* dan *RMSE* untuk semua metode. Dengan demikian, distribusi kemampuan kelompok berpengaruh terhadap kualitas penyetaraan vertikal model kredit parsial. Hasil ini mendukung Nonny Swediati (1997), bahwa keakuratan hasil penyetaraan dipengaruhi oleh perbedaan rata-rata kemampuan kelompok *Target* dan *Base*; dan selaras Kim & Cohen (2002), bahwa kondisi kelompok *Target N*(0,1) dan *N*(1,1) berpengaruh terhadap pencapaian *RMSD*.

Implikasi terhadap Pengukuran Hasil Belajar Siswa

Pencapaian tingkat kemampuan θ terhadap seluruh butir tes secara simultan menunjukkan bahwa rata-rata kemampuan θ tertinggi 0,542 (kelas IX), –0,015 (kelas VIII), dan terendah –0,477 (kelas VII). Setiap pasangan kelas memiliki perbedaan signifikan (p<0,05) (F=166,459; Sig=0; homogenitas varians teruji pada p>0,05). Berkenaan dengan soal Matematika model *IRT* politomus *PCM*, siswa kelas IX memiliki pengalaman belajar Matematika lebih tinggi daripada kedua kelas lain.

Rentang rata-rata kemampuan θ siswa kelas VII, kelas VIII, dan kelas IX secara simultan adalah $-1{,}966 \leq \theta \leq 2{,}798$. Rentangan ini memuat siswa

kelas VII yang memiliki kemampuan θ tertinggi, dan terendah untuk siswa kelas IX. Dengan demikian, sebagian siswa kelas VII memiliki kemampuan setingkat dengan sebagian siswa kelas di atasnya.

Perkembangan hasil belajar siswa kelas VII menuju kelas VIII menunjukkan hubungan linear dengan persamaan: $Y_1 = 0{,}190\ X_1 + 0{,}472$; ($Y_1$: hasil belajar kelas VIII, dan X_1: hasil belajar kelas VII). Dengan kondisi (a) persyaratan normalitas terpenuhi; (b) linearitas regresi teruji pada $p<0{,}05$ ($F=36{,}651$); (c) koefisien regresi 0,190 ($t=6{,}054$) dan konstanta 0,472 ($t=13{,}588$) teruji signifikansi pengaruhnya ($p<0{,}05$); dan (d) homoskedastisitas teruji pada $p>0{,}05$. Perkembangan hasil belajar siswa kelas VIII menuju kelas IX menunjukkan hubungan linear dengan persamaan: $Y_2 = 0{,}281\ X_2 + 0{,}375$; ($Y_2$: hasil belajar kelas VIII, dan X_2: hasil belajar kelas IX). Dengan kondisi (a) persyaratan normalitas terpenuhi; (b) linearitas regresi teruji pada $p<0{,}05$ ($F=59{,}895$); (c) koefisien regresi 0,281 ($t=7{,}739$) dan konstanta 0,375 ($t=9{,}539$) teruji signifikansi pengaruhnya ($p<0{,}05$); dan (c) homoskedastisitas teruji pada $p>0{,}05$.

Dengan demikian, perkembangan hasil belajar siswa kelas VII menuju kelas VIII dan kelas VIII menuju kelas IX, masing-masing membentuk garis linear. Semakin meningkat kemampuan θ bidang Matematika pada kelas lebih rendah, semakin meningkat pula kemampuan θ bidang Matematika pada tingkat kelas di atasnya.

Hasil kalibrasi butir tes *anchor* kelas VII dan VIII (butir b_{13}, b_{14}, b_{15}, b_{21}, b_{24}), kelas VIII dan kelas IX (butir c_{12}, c_{13}, c_{14}, c_{15}, c_{23}) secara simultan menunjukkan rata-rata kemampuan θ tertinggi 0,597 (kelas IX); 0,096 (kelas VIII), dan terendah –0,641 (kelas VII). Dengan *Test Characteristic Curve*, diperoleh harga $P_i(\theta)$ (mendekati) untuk masing-masing θ secara grafis, yaitu (0,597, 0,790); (0,096, 0,565); (-0,641, 0,210). Secara simultan, perkembangan kemampuan θ siswa dapat ditunjukkan secara grafis.

Simpulan dan Saran

Simpulan

Hasil penelitian menunjukkan bahwa:

1. Ukuran sampel berpengaruh terhadap kualitas penyetaraan vertikal model kredit parsial. Keakuratan penyetaraan semakin meningkat seiring dengan semakin mening-katnya ukuran sampel. Ukuran sampel minimum (300) memiliki rata-rata *RMSD* dan *RMSE* cukup kecil sehingga dapat menghasilkan kualitas penyetaraan yang baik.
2. Panjang tes berpengaruh terhadap kualitas penyetaraan vertikal model kredit parsial. Keakuratan penyetaraan semakin meningkat seiring dengan semakin meningkatnya panjang tes. Penyetaraan dengan panjang tes 20 butir hasilnya lebih akurat daripada penyetaraan dengan panjang tes 10 butir untuk semua situasi.
3. Panjang tes *anchor* berpengaruh terhadap kualitas penyetaraan vertikal model kredit parsial. Keakuratan penyetaraan semakin meningkat seiring bertambahnya panjang tes *anchor*. Dengan rentang panjang tes *anchor* 25% – 30% untuk butir politomus, model kredit parsial dengan panjang tes 20 butir menggunakan panjang tes *anchor* minimum 5 butir dan panjang tes 10 butir menggunakan minimum 3 butir.
4. Metode *M/M* cenderung lebih akurat daripada ketiga metode yang lain dalam penyetaraan vertikal *IRT* butir tes Matematika politomus *PCM*. Keakuratan hasil penyetaraan vertikal skor butir *IRT* politomus model kredit parsial, secara berurutan, cenderung ditunjukkan metode *M/M*, *S-L*, *M/S*, dan *HA*.

Saran

Berdasarkan hasil penelitian, dapat disarankan:

1. Keakuratan hasil penyetaraan semakin meningkat seiring dengan penggunaan ukuran sampel yang semakin besar. Penelitian ini melibatkan ukuran sampel 300, 600, dan 1000. Penerapan ukuran sampel minimal 300 menunjukkan hasil penyetaraan akurat. Namun, perlu dikembangkan ukuran sampel lain yang lebih praktis dan efisien.

2. Penyetaraan dengan panjang tes 20 butir hasilnya lebih akurat dibandingkan dengan panjang tes 10 butir untuk butir politomus model kredit parsial. Panjang tes 20 butir politomus memberikan kelonggaran variasi materi, namun perlu penanganan lebih cermat (seperti distribusi materi, kedalaman materi, kebahasaan). Perlu diselidiki panjang tes yang lebih bervariasi agar diperoleh panjang tes yang lebih efektif.
3. Penggunaan rentang panjang tes *anchor* 25% – 30% untuk butir politomus, model kredit parsial memerlukan panjang tes *anchor* minimum 5 butir untuk panjang tes 20 butir dan minimum 3 butir untuk panjang tes 10 butir. Namun, perlu dikembangkan panjang tes *anchor* butir politomus yang lebih variatif sehingga ditemukan panjang tes *anchor* minimum yang hasil aplikasinya lebih akurat.
4. Metode *M/M* menghasilkan *RMSD* dan *RMSE* lebih kecil (secara keseluruhan) daripada ketiga metode lainnya. Pada butir tes model kredit parsial, metode *M/M* memiliki koefisien penyetaraan *slope* senantiasa 1 sehingga *RMSD* selalu nol. Untuk menghindari kondisi demikian, perlu dipertimbangkan penerapan indeks diskrimi-nasi tertentu pada *GPCM* dalam memperoleh *PCM* pada penelitian lebih lanjut.
5. Penyetaraan vertikal model *IRT* politomus *PCM* melibatkan dua kelompok *Target* berdistribusi *N*(0,1) dan *N*(1,1); dan satu kelompok *Base* berdistribusi *N*(0,1) siswa SMP. Penelitian ini dapat dikembangkan pada jenjang sekolah lainnya dengan kelompok *Base* lebih dari satu dan distribusi kemampuan yang lebih bervariasi.

Daftar Pustaka

Adams, R. J., & Khoo, S. T. (1996). *QUEST: The interactive test analysis system*. Camberwell, VA: ACER Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baker, F. B. (1993). Equating test under the nominal response model. *Applied Psycho-logical Measurement, 17*(3), 239–251.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*, 147–162.

Cook, L. L., & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement. Issues and Practice, 10*, 37–45.

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

Dodd, B. G. & de Ayala, R. J. (1994). Item information as a function of threshold values in the rating scale model. Dalam M. Wilson (Ed.), *Objective Measurement: Theory into Practice* (pp. 299-315). Norwood, NJ: Ablex Publishing Corporation.

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

Ferrara, S. & Walker-Bartnick. (29 Maret 1989). *Constructing an essay prompt bank using the partial credit model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. Diambil pada tanggal 18 Maret 2010 dari http://www.education.umd.edu/EDMS/MARCES/mdarch/pdf/M013027.pdf

Gifford, J. A. & Swaminathan, H. (1990). Bias and the effect of priors in bayesian estimation of parameters of item response models. *Applied Psychological Measurement, 14*(1), 33–43.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Iowa Testing Programs Occasional Papers*, 17. Abstract. Diambil pada tanggal 12 Februari 2004, dari http://SearchERIC.org/ericda/ED193300.htm

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). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Han, K. T. & Hambleton, R. K. (2007). *User's manual: WinGen2*. Amherst, MA: University of Massachusetts, Center for Educational Assesment.

Harris, D. J. (1991). *A comparison of Angoff's design I and Angoff's design II for vertical equating using traditional and IRT methodology*. Abstract. Diambil pada tanggal 12 Februari 2004, dari http://SearchERIC.org/ericda/EJ35192.htm

Hoe, S. L. (2008). Issue and procedures in adopting structural equation modeling technique. *Journal of Applied Quantitative Methods*, *3*, 1, 76-83.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological theory*. Homewood, IL: Dow Jones-Irwin.

Jöreskog, K. G. & Sörbom, D. (1996). *LISREL 8*. *User's reference guide*. Chicago: Scientific Software International.

Kennedy, L. M., Tipps, S., & Johnson, A. (2008). *Guiding children's learning of mathematics* (11th ed.). Belmont, CA: Thomson Wadsworth.

Kim, S. H., & Cohen, A. S. (2002). A comparasion of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 26*(1), 25–41.

Kim, S. & Kolen, M. J. (2004). *STUIRT: A computer program for scale transformation under unidimensional item response theory models*. Iowa, IA: The University of Iowa, Iowa Testing Programs.

Kirisci, L, Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumpsions of unidimensionality and normality. *Applied Psychological Measurement, 25*(2), 146–162.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Loyd, B. H., & Hoover, H. D. (1979). *A comparison of methods vertical equating*. Abstract. Diambil pada tanggal 12 Februari 2004, dari http://SearchERIC.org/ericda/ED177199.htm

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica, 47*(2), 149–174.

Nonny Swediati. (1997). *Test equating under generalized partial credit model*. Unpublished Dissertation, University of Massachusetts Amherst.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25*(1), 53–67.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. Dalam R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221–262). New York: American Council on Education, Macmillan Publishing Company.

Stocking, M. L.& Lord, F. M. (1980). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201–210.

Syaifuddin, M. (2005). *Penyetaraan tes model respons berjenjang*. Disertasi doktor, tidak diterbitkan. Yogyakarta: Program Pascasarjana Universitas Negeri Yogyakarta.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. Dalam R. L. Brennan (Ed.), *Educational Measurement* (4th ed. pp. 111–154). Westport, CT: American Council on Education and Praeger Publishers.