

## PENYETARAAN VERTIKAL DENGAN PENDEKATAN KLASIK DAN ITEM RESPONSE THEORY PADA SISWA SEKOLAH DASAR

<sup>1)</sup>Anak Agung Purwa Antara, <sup>2)</sup>Bastari

<sup>1)</sup>IKIP Saraswati Tabanan, <sup>2)</sup>Universitas Indonesia

<sup>1)</sup>purwa.antara@gmail.com, <sup>2)</sup>bastari@kemdikbud.go.id

### Abstrak

Penelitian ini bertujuan untuk mengetahui (1) akurasi metode linear dan ekipersentil pada pendekatan Klasik dan *mean & mean* dan *mean & sigma* pada pendekatan teori respons butir; (2) karakteristik tes prestasi belajar matematika model campuran kelas III, IV, V, dan VI Sekolah Dasar yang digunakan dalam penelitian ini; dan (3) perkembangan kemampuan siswa ( $\theta$ ) kelas III, IV, V, dan VI pada pelajaran matematika berdasarkan tes yang digunakan. Data penelitian adalah skor siswa pada ujian akhir semester tahun 2013 yang dikumpulkan dengan desain tes dengan butir bersama. Hasil penelitian menunjukkan bahwa, (1) *Standart Error of Equating (SEE)* metode ekipersentil 0,505116 dan metode linear 0,608, sedangkan *Root Mean Square Different (RMSD)* metode *mean & sigma* 0,344 dan metode *mean & mean* 0,394, perbedaan nilai *SEE* pada pendekatan Klasik dan *RMSD* pada pendekatan *IRT* sangat kecil sehingga kebermaknaannya masih perlu diteliti; (2) tes prestasi belajar matematika kelas III, IV, V, dan VI berturut-turut memiliki rerata *location* katagori sedang -0,070, -0,228, -0,324, -0,430, *slope* yang baik 0,537, 0,558, 0,639, 0,644, dan *items fit statistics* minimum 0,077, 0,052, 0,055, dan 0,198; (3) keempat metode penyetaraan memberikan informasi bahwa nilai  $\theta$  meningkat seiring dengan meningkatnya peringkat kelas, dengan rerata  $\theta$  kelas III, IV, V, dan VI berturut-turut sebesar -0,104, -0,076, 0,004, dan 0,173. Namun, kenaikan nilai tersebut sangat kecil sehingga kebermaknaannya masih perlu diteliti.

**Kata kunci:** *Metode penyetaraan, Tes model campuran*

## VERTICAL EQUATING USING CLASICAL METHOD AND ITEM RESPONSE THEORY IN ELEMENTARY SCHOOL STUDENTS

<sup>1)</sup>Anak Agung Purwa Antara, <sup>2)</sup>Bastari

<sup>1)</sup>IKIP Saraswati Tabanan, <sup>2)</sup>Universitas Indonesia

<sup>1)</sup>purwa.antara@gmail.com, <sup>2)</sup>bastari@kemdikbud.go.id

### Abstract

The study was aimed at finding: (1) the accurate linear method and equipercntile method in Classic approach and *mean&mean* and *mean & sigma* method in *IRT* approach, (2) the characteristics of the mathematic mixed-model achievement test for elementary students grade III, IV, V, and VI used in this study, (3) the progress of elementary students gade III, IV, V, and VI ability ( $\theta$ ) in Mathematic based on the developed test. The data of this study were students' scorer on final exams Elementary school 2013 collected with anchor test design. The result showed that: (1) Standard Error of Equating (SEE) within the linear method was 0.607 and ekipercntile method was 0.505116, while Root Mean Square Different (RMSD) within the *mean&mean* was 0.394 and *mean&sigma* method was 0.344. The differences between the value of SEE and RMSD was very little, therefore the meaningfulness needed further study; (2) the Mathematic achievement test for Grade III, IV, V, and VI had location in the middle category within the values respectively -0.070, -0.228, -0.324, -0.430, the good mean of slope 0.537, 0.558, 0.639, 0.644, and items fit statistics minimum 0.077, 0.052, 0.055, 0.198; (3) the four equation method provided information that  $\theta$  increased in accordance with the increased of the class rank, with mean of the  $\theta$  students Grade III, IV, V, and VI as 0.104, -0.076, 0.004, dan 0.173. But the increased of the  $\theta$  value was relatively small, in that its meaningfulness also required further study.

**Keywords:** *equating method, mixed-model test*

## **Pendahuluan**

Isu penting yang selalu hangat dibicarakan dalam dunia pendidikan adalah peningkatan kualitas hasil belajar atau kompetensi lulusan dari suatu lembaga pendidikan agar lulusan yang dihasilkan menjadi sumber daya yang bermanfaat, mampu bertahan dan bersaing dalam kehidupan sekarang ataupun dalam kehidupan yang akan datang. Usaha-usaha pencapaian kualitas hasil belajar tersebut telah dilakukan pemerintah misalnya melalui penyempurnaan kurikulum yang mencakup empat hal yaitu Standar Kompetensi Lulusan (SKL), materi yang harus dipelajari atau diajarkan, pelaksanaan atau proses pembelajaran, dan penilaian ketercapaian tujuan pembelajaran atau hasil belajar (Kemendikbud, 2013).

Berkaitan dengan hal tersebut, Mardapi (2008, p.5) menyatakan bahwa, peningkatan kualitas hasil belajar dari suatu lembaga pendidikan dapat dimulai dari peningkatan kualitas program pembelajaran dan peningkatan kualitas penilaian yang dilakukan di dalam kelas. Hal tersebut dapat dimaknai bahwa guru sebagai komponen yang bersentuhan langsung dengan siswa hendaknya mampu mengembangkan potensi-potensi yang dimiliki siswa, baik melalui pelaksanaan program pembelajaran, maupun melalui model penilaian yang digunakan. Model penilaian yang digunakan guru, hendaknya mampu memantau atau memberikan informasi tentang proses pembelajaran yang telah dilakukan, perkembangan kemampuan siswa setelah mengikuti proses pembelajaran, dan hasil belajar yang telah dicapai siswa, sehingga dapat dilakukan evaluasi untuk perbaikan menuju pencapaian tujuan pendidikan yang diharapkan.

Penilaian yang baik memerlukan data yang benar sehingga perlu didukung oleh proses pengukuran yang baik (Mardapi, 2008, p.5). Penerapan program pengukuran yang bermutu akan mendorong semua komponen yang terlibat dalam proses pembelajaran, terutama guru dan siswa untuk melakukan upaya-upaya dalam pencapaian kualitas hasil belajar sesuai dengan tujuan yang diharapkan. Guru dituntut meningkat-

kan pengetahuan dan memperluas penguasaan materi pelajaran sehingga mampu menyusun tes sebagai alat ukur yang berkualitas untuk mengukur ketecapaian program pembelajaran yang telah dilakukan.

Alat ukur pada pelajaran matematika memiliki struktur tersendiri karena materi matematika bersifat hierarkis. Hal ini berarti penguasaan siswa terhadap materi pada kelas tertentu, berkaitan dengan penguasaan materi pada kelas sebelumnya. Dengan demikian, untuk mengungkap perkembangan kemampuan siswa pada pelajaran matematika diperlukan kemampuan menyusun tes yang dapat menghubungkan kemampuan siswa pada seluruh peringkat kelas. Guru hendaknya mampu memilih model tes yang sesuai, mampu menyusun tes yang baik, dan menggunakan tes tersebut secara tepat untuk memperoleh informasi perkembangan kemampuan siswa.

Rasiman (2008, p.258), menyatakan bahwa informasi perkembangan kemampuan siswa dapat diperoleh secara optimal apabila guru memperhatikan perkembangan kognitif siswa yang diukur. Penggunaan tes yang berada di luar kecakapan siswa akan menyebabkan siswa tersebut tidak mampu menjawab butir soal yang diberikan sehingga guru tidak memperoleh informasi yang cukup tentang perkembangan kemampuan siswanya. Perbedaan kemampuan siswa yang ekstrim dalam satu kelas akan menyebabkan guru menemui masalah dalam melakukan pengukuran. Siswa yang pintar kemungkinan akan mampu menjawab secara tuntas semua butir soal yang berada pada tingkatnya, dan ada kemungkinan mampu menjawab soal dengan materi yang berada di atas tingkatnya, demikian sebaliknya.

Menurut Naga (1992, p.347), cara untuk mengatasi masalah tersebut adalah dengan memberikan siswa yang lambat tes yang berada pada tingkat di bawahnya, sedangkan siswa yang cepat diberikan tes pada tingkat di atasnya, kemudian melalui penyamaan skala, skor yang mereka peroleh dikonversi kembali pada kedudukan semula.

Menurut Kolen & Brennan (2004, p. 5), skor dua tes berbeda dari dua kelompok

atau lebih dapat diperbandingkan jika butir-butir soalnya setara dan memiliki skala yang sama. Kesetaraan antarskor dapat dilakukan secara statistik. Proses statistik yang digunakan untuk menghasilkan skala tunggal dari skor dua tes yang berbeda tersebut dinamakan penyetaraan atau *equating* (Kolen & Brennan, 1995, p.2).

Penyetaraan dapat dilakukan secara horizontal ataupun secara vertikal (Crocker & Algina, 1986, p.456; Hambleton & Swaminathan, 1985, p.197). Penyetaraan yang dilakukan terhadap tes yang memiliki tingkat kesulitan berbeda pada tingkat kelas yang berbeda tapi mengukur *trait* yang sama disebut penyetaraan vertikal. Penyetaraan vertikal dapat digunakan guru untuk mengungkap perkembangan kemampuan siswa, walaupun siswa tersebut berada pada tingkat kelas yang berbeda dan memiliki tingkat kemampuan berbeda, asalkan tes yang digunakan mengukur *trait* yang sama.

Penggunaan penyetaraan vertikal pada tingkat kelas yang berbeda pada pelajaran matematika akan membantu guru memperoleh informasi tentang kemampuan siswa pada mata pelajaran tersebut sehingga perkembangan kemampuan siswa tersebut dapat dipantau. Dengan demikian, sangat jelas bahwa betapa pentingnya melakukan penyetaraan vertikal bagi lembaga pendidikan karena melalui penyetaraan vertikal dapat diperoleh informasi tentang kemampuan siswa dan berbagai kecenderungannya. Informasi tersebut akan menjadi acuan dalam usaha perbaikan menuju peningkatan kualitas hasil belajar atau lulusan pada lembaga tersebut.

Penelitian tentang penyetaraan telah banyak dilakukan oleh para peneliti, di antaranya penyetaraan dengan pendekatan *IRT* untuk butir dikotomis dengan metode *mean & mean*, *mean & sigma*, dan *Stocking-Lord* oleh Kim et al. (2009); penyetaraan dengan pendekatan *IRT* menggunakan metode *Stocking-Lord*, *mean & sigma* dan *concurrent calibration* untuk format model campuran oleh Pang et al. (2010); penyetaraan dengan pendekatan Klasik menggunakan metode linear untuk tes model campuran *multiple-*

*choice* dan *constructed-response* oleh Kim, Walker, & Frederick (2010); penyetaraan dengan pendekatan *IRT* menggunakan metode *characteristic curve* dan *mean & sigma* untuk butir *GPCM* oleh Swediati (1997); penyetaraan dengan pendekatan *IRT* menggunakan metode *stocking-lord*, *mean & sigma* pada respon berjenjang oleh Syaifuddin (2005); penyetaraan dengan pendekatan *IRT* menggunakan metode *mean & mean*, *mean & sigma*, *baebara*, *stocking-lord* pada butir *PCM* oleh Sugeng (2010); dan masih banyak penelitian-penelitian lainnya. Penelitian-penelitian tersebut kebanyakan menggunakan pendekatan *IRT* dan masih sedikit menggunakan pendekatan Klasik.

Hasil penelitian yang telah ada menunjukkan, tidak ada suatu metode yang unggul pada semua kondisi. Suatu metode penyetaraan mungkin akurat pada kondisi tertentu tapi tidak akurat pada kondisi yang lain. Oleh karena itu, masih perlu diteliti lebih lanjut keakuratan berbagai metode penyetaraan yang ada baik dengan pendekatan Klasik maupun *IRT*. Guru perlu memperoleh informasi, metode mana yang memberikan hasil penyetaraan yang akurat pada kondisi yang ada, sehingga informasi yang diperoleh dalam memantau perkembangan kemampuan siswanya mendekati kebenaran.

Dari uraian yang telah disampaikan sebelumnya, penelitian ini bertujuan untuk mengetahui: (1) akurasi metode linear dan ekipersentil pada pendekatan Klasik dan *mean & mean* dan *mean & sigma* pada pendekatan *IRT*; (2) karakteristik tes prestasi belajar matematika model campuran kelas III, IV, V, dan VI Sekolah Dasar yang digunakan dalam penelitian ini; dan (3) perkembangan kemampuan siswa ( $\theta$ ) kelas III, IV, V, dan VI pada pelajaran matematika berdasarkan tes yang digunakan.

## Metode Penelitian

Penelitian ini adalah penelitian tentang penyetaraan vertikal yang secara umum menggunakan pendekatan kuantitatif, dan dalam pelaksanaannya terdiri dari beberapa bagian penelitian yang saling terkait yaitu

penyusunan instrumen penelitian, analisis karakteristik instrumen yang disusun, perbandingan keakuratan metode penyetaraan pada pendekatan klasik dan IRT, dan penelitian tentang perkembangan kemampuan siswa.

Penelitian diawali dengan pengembangan instrumen (tes) prestasi belajar matematika model campuran untuk kelas III, IV, V, dan VI Sekolah Dasar yang diujikan pada semester 2 (tes sumatif) yang disusun berdasarkan pokok bahasan bilangan, geometri dan pengukuran sesuai dengan Kurikulum Tingkat Satuan Pendidikan (KTSP). Empat paket tes memuat butir-butir *anchor* sebanyak 20 % baik untuk pilihan ganda maupun uraian. Penentuan Standar Kompetensi, Kompetensi Dasar dan indikator yang diujikan dilakukan melalui *FGD (forum group discussion)*. Peserta *FGD* adalah guru kelas, guru mata pelajaran, ahli bidang studi (dosen matematika), dan penulis. Penyusunan kisi-kisi dan penulisan soal dilakukan oleh tim yang terdiri dari dua orang guru senior mata pelajaran matematika Sekolah Dasar. Validitas isi dan keterbacaan soal melibatkan dua ahli (*expert*) dalam bidang pendidikan matematika dan pengukuran. Validitas isi dianalisis menggunakan teknik Gregory (2000, p.123). Keterbacaan soal melibatkan 6 guru dan 12 orang siswa kelas VI Sekolah Dasar. Revisi soal dilakukan setelah hasil uji keterbacaan dan uji validitas isi dari *expert* diperoleh. Instrumen (tes) yang telah diperbaiki selanjutnya diujicoba di sebelas Sekolah Dasar. Data hasil ujicoba dianalisis menggunakan pendekatan Klasik dan IRT menggunakan program *Parscale* (Muraki & Bock, 1977) dengan estimasi *Marginal Maximum Likelihood (MML)* untuk mengetahui karakteristik butir tes yang disusun sehingga dapat dipilih butir-butir yang baik. Butir-butir yang telah diseleksi pada masing-masing paket tes, selanjutnya dirakit menjadi tes akhir semester yang digunakan sebagai instrumen pengambilan data penelitian.

Pengumpulan data penelitian dilakukan dengan *stratified random sampling* melibatkan sampel sebanyak 1034 terdiri dari 257

siswa kelas III, 257 siswa kelas IV, 260 siswa kelas V, dan 260 siswa kelas VI. Penerapan random dilakukan pada tingkat sekolah, sedangkan penentuan strata sekolah dengan memperhatikan letak sekolah dan katagori sekolah.

Pemenuhan asumsi unidimensi dan validitas konstruk dari tes yang digunakan dilakukan dengan analisis faktor eksploratori dan konfirmatori. Banyaknya dimensi yang diukur oleh keempat tes, dilihat dari *Scree-plot* nilai *Eigen*. Hal ini sesuai dengan pendapat Demars (2010, p. 39) bahwa *eigenvalue* dari inter-item matriks korelasi adalah salah satu metode yang *simple* untuk mengetahui dimensionalitas data. Pengujian kecocokan model hipotetik pengukuran terhadap data empiris menggunakan analisis faktor konfirmatori yang dilakukan menurut peringkat kelas terhadap kelompok butir soal yang terdapat pada tes tersebut. Program yang digunakan adalah Lisrel 8.54 dengan indikator *goodness of fit* (Joreskog & Sorbom, 1996, p. 27) yang meliputi *Non Normed Fit Index (NNFI)*, *Comparative Fit Index (CFI)*, dan *Root Mean Square Error of Approximation (RMSEA)*, dan secara umum dengan statistik  $\chi^2$ . Indikator *NNFI* dan *CFI* dengan kriteria lebih dari 0,90 menunjukkan kecocokan yang baik, dan nilai *RMSEA* kurang dari 0,08 menunjukkan kecocokan yang layak. Statistik  $\chi^2$  dengan rasio  $\chi^2 / df \leq 3$ .

Desain penyetaraan menggunakan *common items non-equivalent groups* desain atau desain tes *anchor* dengan metode linear dan ekipersentil pada pendekatan Klasik dan dengan metode *mean & mean* dan *mean & sigma* pada pendekatan IRT.

Akurasi metode linear dan ekipersentil diuji dengan *Standard Error of Equating (SEE)* yang dihitung dengan rumus (Linn, 1989, p. 255),

$$SEE [Y(X)] = \{[(\sigma^2_Y / 2)] [2 + (X_i - \mu_X)^2 / \sigma^2_X] [(1/N_X) + (1/N_Y)]\}^{1/2}$$

dengan,

*SEE [Y(X)]* adalah standar kesalahan penyetaraan dari tes X ke Y,  $\sigma^2_Y$  adalah varians skor tes Y,  $\sigma^2_X$  adalah varians skor tes X,  $X_i$

adalah skor ke  $i$  tes  $X$ ,  $\mu_X$  adalah rerata skor tes  $X$ ,  $N_X$  adalah jumlah peserta tes  $X$ ,  $N_Y$  adalah jumlah peserta tes  $Y$ .

Metode *mean & mean* dan *mean & sigma* akurasi diuji dengan *Root Mean Square Different (RMSD)* kemampuan siswa ( $\theta$ ) sebelum dan sesudah penyetaraan, yang dihitung dengan rumus (Kim & Cohen, 2002, p. 31),

$$RMSD(\theta) = \sqrt{\frac{\sum_{i=1}^N (\hat{\theta}_i - \theta_i)^2}{N}}$$

dengan,

$N$  adalah banyaknya *examinee*,  $\hat{\theta}_i$  adalah kemampuan peserta ke  $i$  setelah disetarakan, dan  $\theta_i$  adalah kemampuan peserta ke  $i$  sebelum disetarakan.

Informasi tentang perkembangan kemampuan siswa dalam pelajaran matematika dapat diperoleh dari hasil penyetaraan skor dan skala kemampuan siswa pada keempat kelas yang dilibatkan.

### Hasil Penelitian dan Pembahasan

Data penelitian yang berupa skor siswa pada tes prestasi belajar matematika model campuran pada ujian akhir semester kelas III, IV, V, dan IV Sekolah Dasar dikoreksi oleh dua orang *rater* untuk mendapatkan skor yang baik. Untuk menjamin konsistensi penilaian, skor dari dua orang *rater* tersebut diestimasi reliabilitasnya dengan menggunakan formula *inter-rater* dengan pendekatan *Hoyt* (Mardapi, 2012, p. 86).

Koefisien reliabilitas ( $r_{11}$ ) *inter-rater* skor tes kelas III, IV, V, dan IV diperoleh berturut-turut sebesar 0,992, 0,823, 0,923 dan 0,926. Nilai-nilai koefisien tersebut semuanya berada di atas 0,700. Hal tersebut berarti unsur subjektivitas dari masing-masing *rater* (penilai) tidak berpengaruh terhadap pemberian skor pada ke empat tes, sehingga skor yang diberikan dapat digunakan sebagai data penelitian.

### Uji Unidimensi dan Validitas Konstruk

Nilai *Chi-Square* pada uji Bartlet tes kelas III, sebesar 1217,326 dengan derajat kebebasan 595 dan nilai-p kurang dari 0,01. Hasil ini menunjukkan bahwa ukuran sampel sebesar 257 yang digunakan pada penelitian telah mencukupi. *Scree-plot* nilai *Eigen* tes kelas III menunjukkan nilai *Eigen* mulai landai pada faktor ke dua. Hal ini berarti terdapat satu faktor yang dominan pada tes kelas III. Faktor-faktor lain yang berpengaruh berkaitan dengan faktor dominan tersebut. Hasil *running* dengan program Lisrel mendapatkan nilai *chi-square* sebesar 94,63 dengan  $df = 74$  dan  $p = 0,053$ , *Root Mean Square Error Approximation (RMSEA) = 0,053*, *Non-Normed Fit Index (NNFI) = 0,949*, *Comparative Fit Index (CFI) = 0,958* dan  $\chi^2/df = 94,63/74 = 1,28 < 3$  yang berarti model tes kelas III *fit* dengan data.

Nilai *Chi-Square* pada uji Bartlet tes kelas IV sebesar 1300,659 dengan derajat kebebasan 595 dan nilai-p kurang dari 0,01. Hasil ini menunjukkan bahwa ukuran sampel sebesar 257 yang digunakan pada penelitian juga telah cukup. *Scree-plot* nilai *Eigen* tes kelas IV menunjukkan nilai *Eigen* mulai landai pada faktor ke dua. Hasil *running* dengan program Lisrel mendapatkan nilai *chi-square* diperoleh sebesar 94,58 dengan  $df = 74$  dan  $p = 0,054$ , *Root Mean Square Error Approximation (RMSEA) = 0,053*, *Non-Normed Fit Index (NNFI) = 0,953*, *Comparative Fit Index (CFI) = 0,962* dan  $\chi^2/df = 94,58/74 = 1,28 < 3$ . Hal ini menunjukkan model tes kelas IV *fit* dengan data.

Nilai *Chi-Square* pada uji Bartlet tes kelas V sebesar 1733,177 dengan derajat kebebasan 595 dan nilai-p kurang dari 0,01. Hasil ini menunjukkan bahwa ukuran sampel sebesar 260 yang digunakan pada penelitian telah cukup. *Scree-plot* tes kelas V menunjukkan nilai *Eigen* pada tes kelas V mulai landai pada faktor ke dua. Hasil *running* dengan program Lisrel mendapatkan nilai *chi-square* sebesar 93,24 dengan  $df = 74$  dan  $p = 0,065$ , *Root Mean Square Error Approximation (RMSEA) = 0,053*, *Non-*

Normed Fit Index (NNFI) = 0,951, Comparative Fit Index (CFI) = 0,960 dan  $\chi^2/df = 93,24/74 = 1,26 < 3$ . Hal ini menunjukkan model tes kelas V fit dengan data.

Nilai *Chi-Square* pada uji Bartlett tes kelas VI sebesar 1746,443 dengan derajat kebebasan 595 dan nilai-p kurang dari 0,01. Hal ini berarti ukuran sampel sebesar 260 yang digunakan pada penelitian telah cukup. *Scree-plot* tes kelas VI menunjukkan bahwa nilai *Eigen* pada tes kelas VI mulai landai pada faktor ke dua, yang berarti hanya satu faktor yang dominan yaitu kemampuan matematika. Hasil *running* dengan program Lisrel memperoleh nilai *chi-square* sebesar 122,07 dengan  $df = 100$  dan  $p = 0,06613$ , *Root Mean Square Error Approximation (RMSEA)* = 0,045, *Non-Normed Fit Index (NNFI)* = 0,963, *Comparative Fit Index (CFI)* = 0,969 dan  $\chi^2/df = 122,07/100 = 1,22 < 3$ . Hal ini juga berarti model tes kelas VI fit dengan data.

#### Analisis Butir Soal

Analisis butir dengan pendekatan Klasik meliputi korelasi *Pearson* ( $r_p$ ) dan *Polyserial* ( $r_{ps}$ ) yang dibaca pada *output Parscale Fase 1 (PH1)*. Analisis butir tes dengan pendekatan *IRT* meliputi parameter *slope*, *location*, dan *probability* yang dibaca pada *output Parscale Fase 2 (PH2)*. Rangkuman hasil analisis keempat paket tes disajikan pada Tabel 1, Tabel 2, dan Tabel 3.

Tabel 1 menunjukkan, rerata skor dari tes kelas III, IV, V, dan VI berturut-turut

adalah 25,961, 26,770, 27,25 dan 27,660 berarti di atas rerata ideal (24,50). Varians skor keempat tes cukup besar ini berarti distribusi skor cukup besar. Distribusi skor siswa untuk tes kelas III, IV, dan V membentuk *curve* sedikit juling ke kanan, hal ini dapat dilihat dari nilai rerata sedikit lebih besar dari median. Distribusi skor untuk tes kelas VI membentuk *curve* sedikit juling ke kiri, karena nilai reratanya lebih rendah dari median. Hal ini menunjukkan sebagian besar siswa kelas VI mendapat skor tinggi. Nilai rerata koefisien korelasi *Pearson* dan rerata koefisien korelasi *Polyserial* keempat tes juga berada di atas 0,2 yang berarti keempat tes secara umum memiliki daya beda yang baik.

Tabel 1. Analisis Butir Tes Kelas III, IV, V, dan VI dengan Pendekatan Klasik

Komponen	Tes Kls III	Tes Kls IV	Tes Kls V	Tes Kls VI
Mean	25,96	26,770	27,25	27,660
Varians	30,045	31,146	46,127	46,822
St. Dev.	5,481	5,581	6,792	6,843
<i>Skewness</i>	0,431	0,285	-0,001	-0,090
<i>Kurtosis</i>	-0,034	-0,086	-0,585	-0,614
Skor Min	12	13	10	10
Skor Maks	41	45	42	42
Median	25	26	27	28
Rerata ( $r_p$ )	0,298	0,317	0,369	0,373
Rerata ( $r_{ps}$ )	0,367	0,391	0,455	0,460
Panjang Tes	35	35	35	35
Subjek	257	257	260	260

Tabel 2. Analisis Parameter Butir Tes Kelas III dan IV dengan Pendekatan *IRT*

Komponen	Tes Kelas III			Tes Kelas IV		
	<i>Slope</i>	<i>Location</i>	<i>Prob.</i>	<i>Slope</i>	<i>Location</i>	<i>Prob.</i>
Mean	0,537	-0,070	0,412	0,558	-0,228	0,513
St.Dev.	0,234	0,330	0,231	0,250	0,380	0,277
Varians	0,055	0,109	0,055	0,062	0,145	0,077
Min	0,247	-0,951	0,077	0,295	-1,333	0,052
Maks	1,022	0,525	0,915	1,360	0,536	0,956
<i>Skewness</i>	0,784	-0,498	0,388	1,480	-0,880	-0,188
<i>Kurtosis</i>	-0,610	0,099	-0,920	1,956	1,532	-1,092
Median	0,464	-0,033	0,377	0,475	-0,183	0,498
Panjang Tes	35			35		
Responden	257			257		

Tabel 2 dan Tabel 3 menunjukkan, nilai rerata *location* dari keempat tes berturut-turut sebesar -0,070, -0,228, -0,324, dan -0,430, semuanya berada di sekitar titik nol yang berarti keempat tes memiliki tingkat kesukaran butir yang sedang. Nilai rerata parameter *slope* tes kelas III, IV, V dan VI berturut-turut adalah 0,537, 0,558, 0,639, dan 0,644 (>0,2), berarti keempat tes memiliki daya beda butir yang baik. Nilai minimum *probability* semua butir pada keempat

tes >0,05 yang berarti semua butir tes *fit* dengan model.

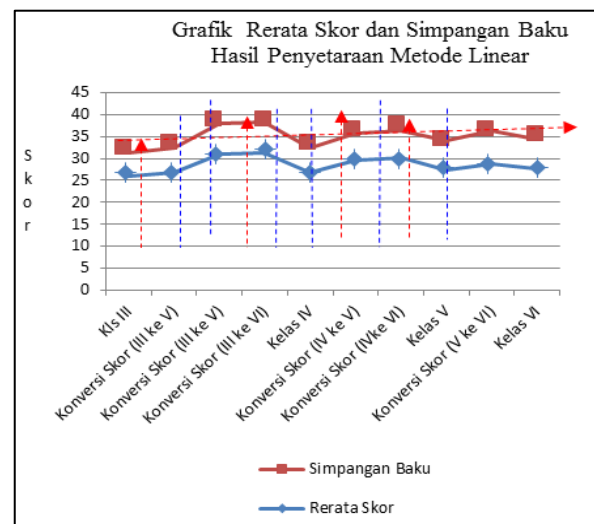
Informasi hubungan antara fungsi informasi tes dengan *Standard Error of Measurement* pada keempat tes memiliki *error* rendah pada rentang skala dari -2 sampai dengan +2, artinya tes akan memberikan informasi yang optimal jika digunakan untuk mengukur kemampuan siswa pada rentang kemampuan antara -2 sampai dengan +2. Hal ini sesuai dengan pendapat Hambleton, Swaminathan, & Rogers (1991, p. 13).

Tabel 3. Hasil Analisis Parameter Butir Tes Kelas V dan VI dengan IRT

Komponen	Tes Kelas V			Tes Kelas VI		
	<i>Slope</i>	<i>Location</i>	<i>Prob.</i>	<i>Slope</i>	<i>Location</i>	<i>Prob.</i>
Rerata	0,639	-0,324	0,417	0,644	-0,430	0,590
St.Dev,	0,310	0,310	0,229	0,317	0,309	0,218
Varians	0,096	0,096	0,053	0,100	0,095	0,047
Min.	0,252	-1,189	0,055	0,255	-1,243	0,198
Maks.	1,447	0,081	0,841	1,502	0,029	0,962
<i>Skewness</i>	1,375	-0,786	0,068	1,364	-0,881	-0,001
<i>Kurtosis</i>	1,216	0,146	-1,147	1,242	0,575	-0,893
Median	0,539	-0,237	0,421	0,533	-0,351	0,586
Panjang Tes	35			35		
Responden	260			260		

### Hasil Penyetaraan

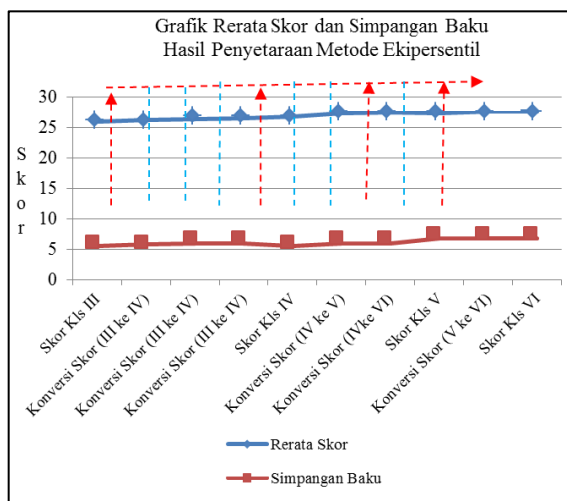
Telah dijelaskan sebelumnya, bahwa penyetaraan vertikal dengan pendekatan klasik menggunakan metode linear dan ekipersentil. Penyetaraan vertikal dengan metode linear mendapatkan persamaan konversi: (1) dari kelas III ke kelas IV adalah  $l_Y(x) = 1,027 X + 0,094$ ; (2) dari kelas III ke kelas V adalah  $l_Y(x) = 1,286 X - 2,014$ ; (3) dari kelas III ke kelas VI adalah  $l_Y(x) = 1,262X - 1,656$ ; (4) dari kelas IV ke kelas V adalah  $l_Y(x) = 1,144 X - 0,553$ ; (5) dari kelas IV ke kelas VI adalah  $l_Y(x) = 1,120 X - 0,411$ ; dan dari kelas V ke kelas VI adalah  $l_Y(x) = 1,098 X - 1,034$ . Rerata skor dan simpangan baku hasil penyetaraan disajikan dengan grafik seperti tampak pada Gambar 1. Besarnya *SEE* penyetaraan dengan metode linear diperoleh sebesar 0,608.



Gambar 1. Hasil Penyetaraan dengan Metode Linear

Penyetaraan skor dengan metode ekipersentil pada dasarnya adalah mendudukkan skor siswa dari keempat tes pada per-

sentil yang sama. Hasil penyetaraan dan simpangan bakunya disajikan dengan grafik pada Gambar 2. Grafik pada Gambar 2 menunjukkan, rerata skor siswa mengalami peningkatan seiring dengan meningkatnya peringkat kelas. Hal ini berarti makin tinggi peringkat kelas kemampuan siswa semakin meningkat. Besarnya nilai *SEE* pada penyetaraan dengan metode ekipersentil diperoleh sebesar 0,505.



Gambar 2. Hasil Penyetaraan Metode Ekipersentil

Penyetaraan dengan pendekatan *IRT* model *GPCM* meliputi penyetaraan parameter daya beda butir atau *slope* ( $a$ ), parameter tingkat kesukaran butir atau *location* ( $b$ ) dan parameter kemampuan siswa atau *ability* ( $\theta$ ). Konstanta penyetaraan  $a$  dan  $\beta$  pada metode *mean & mean* dihitung berdasarkan *mean* dari daya beda butir dan tingkat kesukaran butir, sedangkan pada metode *mean & sigma* konstanta  $a$  dan  $\beta$  dihitung dengan *mean* dan simpangan baku dari parameter *location*.

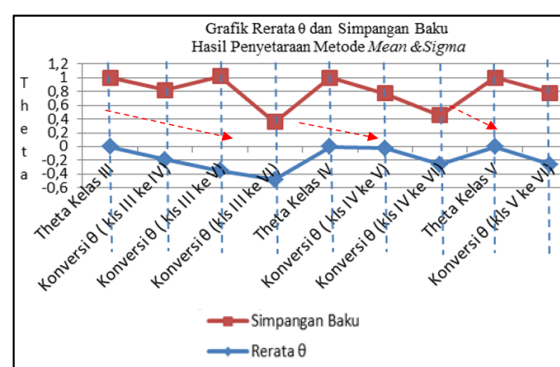
Hasil penyetaraan parameter *slope* dan *location* metode *mean & mean* disajikan pada tabel 4 dan Tabel 5, sedangkan hasil penyetaraan parameter kemampuan  $\theta$  disajikan dengan grafik pada Gambar 3. Akurasi penyetaraan dihitung berdasarkan nilai *RMSD* kemampuan sebelum dan sesudah penyetaraan. Besarnya nilai *RMSD* pada penyetaraan vertikal dengan metode *mean & mean* diperoleh sebesar 0,394.

Tabel 4. Rerata *Slope* dan Simpangan Baku hasil Penyetaraan Tes Kelas III, IV, V, dan VI Metode *Mean & Mean*

Penyetaraan	Rerata <i>Slope</i>	Simpangan Baku
<i>Slope</i> Kls III	0,537	0,234
Konversi <i>Slope</i> (III ke IV)	0,518	0,222
Konversi <i>Slope</i> (III ke V)	0,624	0,277
Konversi <i>Slope</i> (III ke VI)	0,637	0,280
<i>Slope</i> Kls IV	0,558	0,250
Konversi <i>Slope</i> (IV ke V)	0,753	0,367
Konversi <i>Slope</i> (IV ke VI)	0,799	0,365
<i>Slope</i> Kelas V	0,639	0,310
Konversi <i>Slope</i> (V ke VI)	0,672	0,326
<i>Slope</i> Kls VI	0,644	0,316

Tabel 5. Rerata *Location* dan Simpangan Baku Tes Kelas III, IV, V, dan VI Hasil Penyetaraan Metode *Mean & Mean*

Penyetaraan	Rerata <i>location</i>	Simpangan Baku
Tes Kelas III	-0,070	0,330
Konversi Loc. (III ke IV)	-0,267	0,338
Konversi Loc. (III ke V)	-0,482	0,286
Konversi Loc. (III ke VI)	-0,543	0,272
Tes Kelas IV	-0,228	0,380
Konversi Loc. (IV ke V)	-0,207	0,180
Konversi Loc. (IV ke VI)	-0,416	0,254
Tes Kelas V	-0,324	0,310
Konversi Loc. (V ke VI)	-0,585	0,296
Tes Kelas VI	-0,430	0,305



Gambar 3. Grafik Konversi Kemampuan Siswa antar-Peringkat Kelas Metode *Mean & Mean*

Hasil penyetaraan vertikal parameter *location* dan *slope* pada metode *mean & sigma* disajikan pada Tabel 6 dan Tabel 7, sedang-



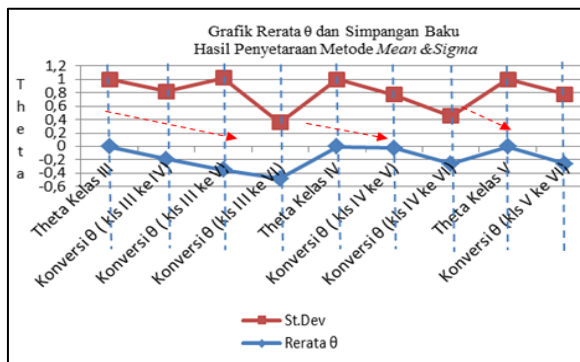
kan hasil penyetaraan parameter kemampuan  $\theta$  disajikan dengan grafik pada Gambar 4. Besarnya nilai *RMSD* metode *mean & sigma* adalah 0,344474.

Tabel 6. Rerata *Location* dan Simpangan Baku Tes Kelas III, IV, V, dan VI Hasil Penyetaraan Metode *Mean & Sigma*

Penyetaraan	Rerata <i>Location</i> (b)	Simpangan Baku
<i>Location</i> Kls III	-0,070	0,330
Konversi <i>Location</i> (III-IV)	-0,262	0,330
Konversi <i>Location</i> (III-V)	-0,448	0,442
Konversi <i>Location</i> (III-VI)	-0,543	0,272
<i>Location</i> Kls IV	-0,207	0,294
Konversi <i>Location</i> (IV-V)	-0,228	0,380
Konversi <i>Location</i> (IV-VI)	-0,390	0,297
<i>Location</i> Kls V	-0,324	0,310
Konversi <i>Location</i> (V-VI)	-0,595	0,321
<i>Location</i> Kls VI	-0,430	0,308

Tabel 7. Rerata *Slope* dan Simpangan Baku Tes Kelas III, IV, V, dan VI Hasil Penyetaraan Metode *Mean & Sigma*

Penyetaraan	Rerata <i>Slope</i> (a)	Simpangan Baku
<i>Slope</i> Kls III	0,537	0,234
Konv. <i>Slope</i> (III-IV)	0,682	0,442
Konv. <i>Slope</i> (III-V)	0,416	0,188
Konv. <i>Slope</i> (III-VI)	0,637	0,280
<i>Slope</i> Kls IV	0,558	0,250
Konv. <i>Slope</i> (IV-V)	0,738	0,307
Konv. <i>Slope</i> (IV-VI)	0,790	0,361
<i>Slope</i> Kls V	0,639	0,310
Konv. <i>Slope</i> (V-VI)	0,622	0,298
<i>Slope</i> Kls VI	0,638	0,317



Gambar 4. Grafik Rerata  $\theta$  dan Simpangan Baku Hasil Penyetaraan Metode *Mean & Sigma*

### Pembahasan

Pada penyetaraan dengan metode linier (Gambar 1) tampak bahwa rerata skor siswa akan naik jika dikonversi ke tingkat kelas di atasnya. Hal tersebut juga berarti bahwa kemampuan siswa mengalami peningkatan seiring dengan meningkatnya peringkat kelas. Namun, jika dicermati persamaan konversi dari kelas III ke V, kelas III ke VI, kelas IV ke V, kelas IV ke VI dan kelas V ke VI maka dapat dilihat siswa yang memiliki skor rendah dan skor tinggi mempunyai arah konversi yang berbeda. Misalnya konversi skor kelas III ke kelas VI memberi informasi bahwa siswa yang mendapat skor 6 ke bawah pada kelas III akan menjadi lebih rendah jika skor tersebut dikonversi ke kelas VI, sedangkan siswa yang mendapat skor 7 ke atas, skornya akan menjadi lebih besar. Persoalan yang sama bisa diselidiki pada persamaan konversi yang lain. Hal ini menunjukkan ketidakkonsistenan terhadap hasil konversi dengan menggunakan metode linear.

Penyetaraan skor dengan metode ekipersentil memberikan informasi yang lebih konsisten dibandingkan metode linear. Hal tersebut dapat dilihat dari grafik rerata skor hasil konversi yang semuanya mempunyai kecenderungan yang sama setelah dilakukan penyetaraan. Rerata skor kelas VI lebih tinggi dari rerata skor kelas V, lebih tinggi dari rerata skor kelas IV, dan lebih tinggi dari rerata skor kelas III, hal yang sama juga berlaku pada kelas V dan IV. Hal ini dapat diartikan bahwa kemampuan siswa mengalami peningkatan sesuai dengan meningkatnya peringkat kelas. Hal tersebut tentu sangat wajar, karena pengetahuan siswa semestinya mengalami peningkatan sesuai dengan pengalaman belajarnya.

Penyetaraan dengan pendekatan *IRT* adalah penyetaraan skala parameter butir suatu tes terhadap skala parameter butir tes lainnya. Hasil penelitian menunjukkan bahwa konversi nilai parameter *location* atau tingkat kesukaran butir menunjukkan hasil yang konsisten baik pada metode *mean & mean* maupun metode *mean & sigma*. Informasi tersebut dapat dilihat pada grafik per-

bandingan nilai parameter *location* sebelum dan sesudah penyetaraan. Nilai rerata parameter *location* menurun jika dikonversi ke skala peringkat di atasnya. Hal ini dapat diartikan bahwa tes kelas III menjadi lebih mudah jika dikerjakan oleh siswa kelas IV, V atau VI. Tes kelas IV menjadi lebih mudah jika dikerjakan oleh siswa kelas V atau VI dan tes kelas V akan lebih mudah jika dikerjakan oleh siswa kelas VI. Sebaliknya, tes kelas VI akan lebih sulit jika dikerjakan oleh siswa kelas V, IV atau III dan seterusnya.

Pada umumnya *slope* dari keempat tes tidak jauh berubah jika dikonversi pada skala tes yang lainnya. Pada metode *mean & mean* indeks daya beda tes kelas III sedikit menurun jika dikonversi ke kelas IV, sedikit lebih baik jika dikonversi ke kelas V atau VI. Indeks daya beda tes kelas IV sedikit lebih baik jika dikonversi ke kelas V, atau VI. Demikian juga tes kelas V daya bedanya sedikit lebih baik jika dikonversi ke kelas VI. Pada metode *mean & sigma*, indeks daya beda tes kelas III sedikit lebih baik jika dikonversi ke kelas IV, sedikit lebih menurun jika dikonversi ke kelas V dan sedikit lebih baik jika dikonversi ke kelas VI. Indeks daya beda tes kelas IV sedikit lebih baik jika dikonversi ke kelas V maupun kelas VI. Indeks daya beda tes kelas V sedikit menurun jika dikonversi ke kelas VI. Secara umum rerata nilai parameter daya beda dari keempat tes mempunyai nilai rata-rata relatif sama. Rerata *slope* tes kelas III, IV, V, dan VI setelah konversi berturut-turut sebesar 0,545, 0,554, 0,523 dan 0,557.

Persamaan konversi kemampuan ( $\theta$ ) mengikuti persamaan ciri butir. Dijelaskan di atas baik pada metode *mean & mean* maupun pada metode *mean & sigma* kemampuan siswa akan menurun jika dikonversi pada skala kemampuan di atasnya. Informasi tersebut dapat dilihat dari grafik konversi rerata kemampuan siswa yang selalu menurun jika dikonversi pada skala di atasnya. Informasi tersebut juga dapat diartikan bahwa rerata kemampuan siswa pada kelas VI lebih tinggi dari rerata kemampuan siswa kelas V. Rerata kemampuan siswa kelas V lebih ting-

gi dari rerata kemampuan siswa kelas IV dan rerata kemampuan siswa kelas IV lebih tinggi dari rerata kemampuan siswa kelas III. Hal ini diperkuat dengan perbandingan rerata kemampuan siswa hasil kalibrasi *concurrent* yang menunjukkan besarnya rerata kemampuan siswa kelas III, IV, V, dan VI berturut-turut sebesar: -0,104, -0,076, 0,004, dan 0,004.

Hasil perhitungan *SEE* pada metode linear mendapatkan besarnya kesalahan standar penyetaraan (*SEE*) sebesar 0,608 sedangkan penyetaraan dengan metode ekipersentil didapat kesalahan standard penyetaraan (*SEE*) sebesar 0,505. Perbedaan nilai *SEE* sebesar 0,102 tidak begitu besar sehingga kebermaknaannya untuk membedakan akurasi kedua masih perlu diteliti.

Hasil perhitungan *RMSD* pada metode *mean & smean* nilai *RMSD* sebesar 0,394 sedangkan penghitungan *RMSD* pada penyetaraan dengan metode *mean & sigma* mendapatkan nilai sebesar 0,344. Kedua nilai *RMSD* tersebut juga tidak jauh berbeda (hanya berselisih 0,049). *RMSD* metode *mean & sigma* sedikit lebih kecil dibandingkan metode *mean & mean*. Kebermaknaan perbedaan nilai *RMSD* untuk membedakan akurasi kedua metode tersebut masih perlu diteliti lebih lanjut melalui penelitian simulasi.

## **Simpulan dan Saran**

### **Simpulan**

Penyetaraan vertikal dengan pendekatan Klasik memberikan informasi bahwa *Standart Error of Equating (SEE)* pada penyetaraan vertikal dengan metode ekipersentil sebesar 0,505 dan *SEE* penyetaraan vertikal dengan metode linear sebesar 0,608. Metode ekipersentil menghasilkan *SEE* lebih kecil dibandingkan metode linear, namun perbedaan nilai *SEE* tersebut sangat kecil sehingga tidak bermakna dalam membedakan akurasi metode linear dan ekipersentil.

Penyetaraan vertikal dengan pendekatan *IRT* memberikan informasi bahwa *Root Mean Square Different (RMSD)* kemampuan ( $\theta$ ) pada penyetaraan vertikal dengan

metode *mean & sigma* sebesar 0,344 dan Nilai *RMSD* kemampuan pada penyetaraan vertikal dengan metode *mean & mean* sebesar 0,394. Nilai *RMSD* yang dihasilkan metode *mean & sigma* lebih kecil dibandingkan metode *mean & mean*, namun, perbedaan nilai *RMSD* tersebut juga sangat kecil sehingga tidak bermakna dalam membedakan akurasi kedua metode.

Penelitian ini menghasilkan empat perangkat tes yaitu tes prestasi belajar matematika semester dua Sekolah Dasar untuk kelas III, kelas IV, kelas V, dan kelas VI yang telah memenuhi syarat sebagai tes yang baik setelah melalui analisis kualitatif maupun kuantitatif. Keempat tes memiliki nilai *location (b)* atau tingkat kesukaran butir kategori sedang ( $-2 < b < 2$ ) dengan nilai *location* berturut-turut sebesar -0,228, -0,324, -0,430, dan -0,430, memiliki daya beda atau *slope (a)* yang baik ( $a > 0,2$ ) dengan nilai rerata *slope* berturut-turut sebesar 0,537, 0,558, 0,639, dan 0,644, serta memiliki *items fit statistics (p)* yang baik ( $p > 0,05$ ) dengan nilai *p* minimum berturut-turut sebesar 0,077, 0,052, 0,055, dan 0,198.

Keempat metode penyetaraan yang digunakan dalam penelitian ini yaitu metode linear, metode ekipersentil, metode *mean & mean*, dan metode *mean & sigma* menunjukkan bahwa rerata kemampuan siswa mengalami peningkatan seiring dengan meningkatnya peringkat kelas. Hal tersebut diperkuat oleh hasil kalibrasi *concurrent* yang menunjukkan perbandingan rerata kemampuan siswa kelas III, IV, V, dan VI mengalami peningkatan setelah dilakukan penyetaraan dengan nilai  $\theta$  berturut-turut sebesar -0,104, -0,076, 0,004, dan 0,173. Namun, kenaikan nilai  $\theta$  dari suatu peringkat ke peringkat di atasnya sangat kecil sehingga masih perlu diteliti kebermaknaannya. Kenaikan tertinggi terjadi pada kelas VI yang disebabkan oleh faktor latihan menjelang Ujian Sekolah.

#### Saran

Hasil penelitian ini memberikan sumbangan pikiran berkaitan dengan kualitas penyetaraan vertikal yang menggunakan dua pendekatan yaitu pendekatan Klasik dan pendekatan *IRT*. Penyetaraan dengan pen-

dekatan Klasik yang membandingkan metode linear dan ekipersentil menunjukkan hasil yang tidak terlalu jauh berbeda. Demikian juga penyetaraan dengan pendekatan *IRT*, yang membandingkan akurasi metode *mean & mean* dan *mean & sigma* hasilnya tidak terlalu jauh berbeda. Perlu dilakukan penelitian simulasi untuk mengetahui signifikansi perbedaan akurasi metode-metode tersebut.

Penyetaraan vertikal dengan pendekatan Klasik dan *IRT* dapat dipertimbangkan penggunaannya dalam tes model campuran. Hal itu dapat digunakan untuk mengamati pertumbuhan kemajuan belajar siswa dari tahun ke tahun (sesuai periode pengukuran), mengamati perbandingan tingkat kemampuan siswa, mengamati perbandingan tingkat kesukaran butir melalui kelompok butir *anchor* menurut jenjang kelas berbeda khususnya bidang studi matematika dan IPA.

Keempat metode penyetaraan memberikan informasi bahwa peningkatan kemampuan siswa dari satu tingkat ke tingkat di atasnya tidak besar sehingga kemungkinan hanya disebabkan oleh faktor umur dan pengalaman belajar. Oleh karena itu, perlu dilakukan perbaikan kualitas program pembelajaran untuk memperoleh peningkatan yang lebih baik.

Penyetaraan vertikal dengan pendekatan Klasik dan *IRT* dapat dipertimbangkan dalam pengembangan bank soal matematika. Hal itu sekaligus sebagai pelapis keamanan dalam pelaksanaan tes.

Penerapan metode linear, ekipersentil, *mean & mean* dan *mean & sigma* memberikan pertimbangan pemilihan metode yang lebih baik. Pemilihan metode melibatkan berbagai program untuk menunjang analisis data, ke-tersediaan dan penguasaan aplikasi program sangat diperlukan.

#### Daftar Pustaka

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

- DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML estimation. *Applied Measurement in Education*, 15, 15-31.
- Gregory, R.J. (2007). *Psychological testing. History, principles, and applications* (4<sup>th</sup> ed). Boston: Pearson Education, Inc.
- Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory*. Boston, MA: Kluwer Inc.
- Hambleton, R.K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage Publication Inc.
- Joreskog, K.G. & Sorbom, D. (1996). *LISREL 8: Structural Equation Modeling*. Chicago: Scientific Software International.
- Kemendikbud. (2013). *Kurikulum 2013. Kompetensi dasar (SD)/madrasah ibtidaiyah (MI)*
- Kim, S.H., & Cohen, A.S. (2002). A comparison of linking and concurrent calibration under the grade response model. *Applied Psychological Measurement*, 26, 25-61.
- Kim, S., & Hanson, B.A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement*, 26, 255-270.
- Kim, S., Walker, M.E., & Frederick, M. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement*, 47, 36-53.
- Kim, J. et. al, (2009, April). *Investigation of vertical scaling using the rasch model*. Paper, presented at the annual meeting of the National Council on Measurement in Education.
- Kolen, M.J., & Brennan, R.L (1995). *Test equating: Methods and practices*. New York: Springer-verlag.
- Mardapi, D. (2008). *Teknik penyusunan tes dan nontes*. Yogyakarta: Mitra Cendikia
- Mardapi, D. (2012). *Pengukuran, penilaian, & evaluasi pendidikan*. Yogyakarta: Nuha Medika
- Muraki, E., & Bock, R.D. (1997). *Parscale : IRT item analysis and test scoring for rating-scale data*. Chicago: Scientific Software International.
- Naga. D.S. (1992). *Pengantar teori skor pada pengukuran pendidikan*. Jakarta: Guna-darma.
- Pang, X., et.al, (2010). A comparison of four test equating methods. *Report Prepared*. Ontario: Education Quality and Accountability office
- Rasiman. (2008, November). *Pengembangan bahan ajar berdasarkan perkembangan kognitif untuk meningkatkan hasil belajar matematika siswa SD*. Makalah disajikan dalam Seminar Nasional Matematika dan Pendidikan Matematika di Universitas Negeri Yogyakarta
- Sugeng. (2010). *Penyetaraan vertikal model kredit parsial soal matematika SMP*. Disertasi Doktor, tidak diterbitkan. Program Pascasarjana, UNY Yogyakarta.
- Swediati, N. (1997). *Equating test under the GPCM*. Disertasi Doktor, tidak diterbitkan. University of Massachusetts, Amherst.
- Syaipuddin. (2005). *Penyetaraan tes model respons berjenjang*. Disertasi Doktor, tidak diterbitkan. Yogyakarta: Program Pascasarjana Universitas Negeri Yogyakarta.