Penyetaraan Tes Berbentuk Uraian

Kartono
Jurusan Matematika FMIPA UNNES

Abstrak

Skor dari dua paket tes yang mengukur kemampuan yang sama yang dibuat dengan kisi-kisi yang sama tidak dapat diperbandingkan langsung, karena skor tes tersebut belum berada pada skala yang sama. Skor tes yang dapat dibandingkan langsung harus terletak pada satu skala, sehingga perlu dilakukan penyetaraan skor tes.Terdapat beberapa metode yang dapat digunakan untuk menyetarakan tes, termasuk tes berbentuk uraian. Masing-masing metode memiliki kekurangan dan kelebihan. Penelitian ini bertujuan untuk menentukan metode penyetaraan tes yang paling stabil pada tes berbentuk uraian. Data pada penelitian ini adalah respon siswa terhadap suatu tes. Ada dua set data, respon siswa kelompok tertentu terhadap tes 1 dan respon siswa kelompok lain terhadap tes 2. Masing-masing set data dianalisis dengan menggunakan program PARSCALE, kemudian konstanta penyetaraan dihitung dengan menggunakan empat metode yaitu metode Rerata & Sigma (RS), Rerata & Rerata (RR), Haebara (HA), dan Stocking & Lord (SL).Semua metode komputasinya menggunakan program STUIRT. Dengan menghitung rata-rata *root mean square differences* (RMSD) kemampuan pada masing-masing metode menurut banyaknya replikasi, nilai rata-rata RMSD yang lebih kecil menunjukkan hasil penyetaraan lebih stabil. Hasil penelitian menunjukkan bahwa nilai rata-rata RMSD kemampuan yang terkecil diperoleh dari metode SL, diikuti metode HA, dan kemudian dua metode lainnya. Diantara keempat metode penyetaraan tes, yang paling stabil adalah metode SL. Sebaiknya gunakanlah metode SL dalam penyetaraan tes berbentuk uraian, walaupun komputasinya sukar dilakukan dengan cara manual.
Kata kunci: penyetaraan, PARSCALE, STUIRT, RMSD.

A. PENDAHULUAN

Penggunaan format tes berbentuk uraian dalam penilaian amat populer dan diselenggarakan dalam skala besar, bertarap lokal dan nasional. Tes- tes yang diselenggarakan dalam skala besar untuk kepentingan tertentu biasanya dibuat lebih dari satu paket. Hal ini menunjukkan adanya beberapa paket tes yang digunakan untuk mengukur variabel yang sama, namun skor hasil tes tidak dapat diperbandingkan langsung, karena tes tersebut dibuat pada skala yang berbeda.

Dengan diberlakukannya otonomi sekolah, merupakan kesempatan bagi sekolah untuk menyelenggarakan ujian sendiri sehingga paket tes dengan kisi-kisi yang sama yang diberikan antar sekolah berbeda, sehingga hasilnya tidak

Dipresentasikan dalam Seminar Nasional Matematika dan Pendidikan Matematika 2006 dengan tema **" Trend Penelitian dan Pembelajaran Matematika di Era ICT "** yang diselenggarakan pada tanggal 24 Nopember 2006

dapat dibandingkan langsung karena tes tersebut dibuat pada skala yang berbeda.. Untuk itu paket tes yang beragam untuk mengukur variabel yang sama harus dilakukan penyesuaian terhadap skor-skor tes dalam suatu skala yang sama, sehingga skor pada paket tes yang satu dapat diperbandingkan dengan skor pada paket tes yang lain. Proses statistik yang digunakan untuk menyesuaikan skor-skor tersebut disebut penyetaraan (Kolen, & Brennan, 1995: 2). Dengan penyetaraan tes, tidak hanya skor peserta yang dapat disetarakan, tetapi parameter butir tespun dapat disetarakan.

Suatu penyetaraan tes secara ideal memerlukan syarat–syarat teoretis yang sangat ketat, namun dalam praktik tidak pernah terjadi suatu penyetaraan yang ideal (Kolen, & Brennan, 1995: 246). Syarat-syarat teoretis antara lain menyangkut desain dan metode penyetaraan. Hal ini memiliki pengaruh yang sangat besar pada hasil penyetaraan, disamping faktor lainnya.. Oleh karena itu, untuk meminimalkan ketidakstabilan hasil penyetaraan tes, perlu pemilihan desain dan metode penyetaraan yang tepat.

Ada beberapa metode penyetaraan tes yang dapat digunakan, tidak mungkin untuk membandingkan semua metode yang ada, yang masing-masing memiliki kelebihan dan kekurangan. Oleh karena itu perlu pemilihan metode yang stabil dalam penyetaraan tes. Dalam penelitian ini untuk memilih metode penyetaraan yang stabil, terpilih empat metode yang dibandingkan yaitu metode RS, RR, HA, dan SL. Dua metode yang pertama tersebut disamping formulanya sederhana juga mudah penerapannya, sedangkan dua metode lainnya . walaupun penerapannya tidak semudah kedua metode lainnya, namun kedua metode tersebut dianggap lebih stabil. Diantara keempat metode tersebut, metode manakah yang paling stabil dalam penyetaraan tes berbentuk uraian? Faktor lain yang mempengaruhi kestabilan hasil penyetaraan, adalah ukuran sampel. Ukuran sampel mempengaruhi kestabilan hasil estimasi parameter butir**.**

B. TINJAUAN PUSTAKA

Telah diutarakan bahwa tujuan penyetaraan tes adalah untuk membuat skor pada paket tes yang berbeda tetapi mengukur kemampuan yang sama menjadi setara. Pada bagian ini akan dikaji hal-hal yang berkaitan dengan penyetaraan tes utamanya terfokus pada metode penyetaraan tes berbentuk uraian.

1. Tes Prestasi Belajar

Tes adalah himpunan pertanyaan yang harus dijawab, atau pernyataan-pernyataan yang harus dipilih/ditanggapi, atau tugas-tugas yang dilakukan oleh orang yang di tes (testee) dengan tujuan untuk mengukur suatu aspek (perilaku) tertentu dari testee (Umar, et al., 1997: 7). Tes diklasifikasikan menjadi beberapa macam, tergantung dari tujuannya. Tes yang diberikan dengan tujuan untuk mengukur tingkat penguasaan materi palajaran yang diberikan termasuk tes prestasi belajar. Tes prestasi belajar merupakan suatu

bentuk tes untuk mendapatkan data, yang merupakan informasi untuk melihat seberapa banyak pengetahuan yang telah dimiliki dan dikuasai oleh seseorang sebagai akibat dari pendidikan dan pelatihan (Anastasi & Urbina, 1997: 42).

Tes prestasi belajar, mengukur tingkat kemampuan siswa dalam menguasai bahan pelajaran yang telah diajarkan kepadanya. Tingkat kemampuan yang dimaksud adalah prestasi belajar yang bersifat pengetahuan saja. Berdasarkan bentuknya, tes prestasi belajar dapat dikelompokkan menjadi dua jenis yaitu objektif dan uraian (Gronlund, 1996: 144). Tes bentuk objektif terdiri dari bentuk jawaban singkat, benar salah, menjodohkan, dan objektif uraian dengan lebih dari dua alternatif jawaban..

2. Teori Respon Butir

Ada dua pendekatan dalam mengembangkan pengukuran pendidikan, yaitu pendekatan teori tes klasik dan teori tes modern. Dalam perkembangannya teori tes modern atau disebut pula teori respon butir lebih pesat dari pada teori tes klasik. Hal ini disebabkan oleh adanya keterbatasan pada teori tes klasik. Teori tes klasik mempunyai keterbatasan , bahwa ia bersifat bergantung kelompok dan bergantung butir. Namun demikian, keterbatasan yang ada pada teori tes klasik ini tidak menjadikan surut dalam penerapannya. Terbukti lembaga-lembaga testing yang terkenal, misalnya *ETS (Educational Testing Service), ACT (American Collage Testing Program), NFER (National Foundations for Educational Research), Psychological Corporation* masih menggunakan dasar teori tes klasik dalam mengembangkan tes (Suryabrata, 1998: 35). Hal ini sesuai dengan pendapat yang mengatakan bahwa sebenarnya pendekatan teori tes klasik dapat memberikan informasi mengenai karakteristik butir yang cukup memadai asal dilakukan dengan seksama (Green, Yen, & Burket, 1989: 300). Kelemahan pengukuran semacam ini, tidak terdapat dalam teori respon butir. Teori respon butir memperbaiki keterbatasan yang ada dalam teori tes klasik.

Tujuan teori respons butir ialah membentuk parameter butir dan kemampuan yang bersifat invarians. Invarians parameter butir, artinya butir tes tidak tergantung pada distribusi kemampuan peserta tes dan parameter kemampuan peserta tes tidak bergantung pada parameter butir tes. Kemampuan peserta tes tidak akan berubah hanya karena mengerjakan tes yang berbeda tingkat kesulitannya dan parameter butir tes tidak akan berubah hanya karena diujikan pada kelompok peserta tes yang berbeda kemampuannya.

Invarians parameter butir dapat diselidiki dengan mengujikan tes pada kelompok peserta yang berbeda. Invarians parameter butir terbukti jika estimasi parameter butir tidak berbeda walaupun diujikan pada kelopok peserta yang berbeda tingkat kemampuannya. Invarians parameter kemampuan dapat diselidiki dengan mengajukan dua perangkat tes atau lebih yang memiliki tingkat kesukaran yang berbeda pada kelompok peserta tes. Invarians parameter kemampuan akan terbukti jika estimasi kemampuan peserta tes tidak berbeda walaupun tes yang dikerjakan berbeda tingkat kesulitannya. Sifat invarians tersebut akan terwujud jika ada kecocokan antara perangkat data tes dengan model yang digunakan. Model yang digunakan akan berlaku atau cocok jika data tes memenuhi asumsi-asumsi dalam teori respon butir.

a. Asumsi-asumsi Teori Respon Butir

Dalam teori respon butir, model responnya mempunyai makna bahwa probabilitas subyek untuk menjawab butir dengan benar tergantung pada kemampuan subyek dan karakteristik butir. Hal ini berarti paserta tes dengan kemampuan tinggi akan memepunyai probabilitas menjawab benar lebih besar jika dibandingkan dengan peserta yang mempunyai kemampuan rendah. Hambleton & Swaminathan (1985: 16) dan Hambleton, Swaminathan & Rogers

(1991: 9) menyatakan bahwa ada dua asumsi yang mendasari teori respon butir, yaitu unidimensi, dan independensi local.

Unidimensi, artinya setiap tes hanya berisi satu kemampuan yang diukur oleh butir-butir penyusun tes tersebut. Pada praktiknya asumsi unidimensi tidak dapat dipenuhi secara ketat karena adanya faktor-faktor kognitif, kepribadian dan faktor-faktor administrative dalam tes, seperti kecemasan, motivasi, dan tendensi untuk menebak. Asumsi unidimensi dapat dibuktikan jika tes mengandung hanya satu komponen dominan yang mengukur performansi suatu subyek.

Independensi lokal, terjadi jika kemampuan-kemampuan yang mempengaruhi performansi dijadikan konstan, maka respons subyek terhadap pasangan butir yang manapun akan independen secara statistik satu sama lain. Asumsi ini akan terpenuhi apabila jawaban peserta terhadap sebuah butir soal tidak mempengaruhi jawaban peserta terhadap butir soal yang lain.

Dalam teori respons butir, selain asumsi-asumsi yang harus dipenuhi, hal penting yang perlu diperhatikan adalah pemilihan model yang tepat. Pemilihan model yang tepat akan mengungkap keadaan yang sesungguhnya dari data tes sebagai hasil pengukuran. Ada beberapa model dalam teori respon butir yang dapat digunakan untuk menganlisis butir tes, tergantung bentuk tes.

Pada dasarnya model teori respon butir politomos dapat dikategorikan kedalam dua macam, yaitu model respon nominal dan ordinal, tergantung pada asumsi tentang karakteristik data. Model respon nominal dapat diterapkan pada butir yang mempunyai alternatif jawaban yang tidak terurut dengan adanya berbagai tingkat kemampuan yang diukur, sedangkan model respon ordinal 8digunakan bila respon peserta pada sebuah butir dapat disekor dalam banyaknya kategori tertentu yang tersusun dalam kecakapan. Sebagai contoh, butir-butir skala sikap tipe Likert yang diskor menggunakan sebuah set kategori respon terurut, merupakan penskoran ordinal, butir-butir tes

matematika, fisika, atau kimia dapat diskor dengan menggunakan sistem kredit parsial, dimana poin-poin untuk melengkapi langkah menuju jawaban benar dihargai juga merupakan penskoran ordinal. Diantara model-model politomos tersebut, model-model yang sering digunakan oleh para psikolog antara lain: *Graded Response Models* (GRM) dan *Genelized Partial Credit Models* (GPCM).

3. Penyetaraan Tes

Penyetaraan adalah proses statistik yang digunakan untuk mengatur skor pada format-format tes sehingga skor pada format tersebut dapat diperbandingkan (Kolen , & Brennan, 1995: 2; Yin, Brennan, & Kolen, 2004: 275). Tujuan penyetaraan skor tes adalah untuk mendapatkan skor yang dapat diperbandingkan. Suatu skor dapat dibandingkan dengan skor yang lain, jika keduanya mengukur karakteristik yang sama dan dinyatakan dalam metrik yang sama (Dorans, 2004: 228). Dengan demikian, walaupun kedua skor itu mengukur karakteristik yang sama tetapi kalau skalanya berbeda, kedua skor tidak dapat dibandingkan.

Di dalam penyetaraan tes, perlu diperhatikan sifat-sifat penyetaraan, yaitu sifat setara, simetri, dan grup invarian. Menurut Lord (Hambleton, & Swaminathan, 1985: 199), mengatakan bahwa suatu penyetaraan tes memenuhi sifat setara, jika kedua tes adil untuk setiap peserta, artinya penyetaraan tidak mengenal perbedaan bagi para peserta pada tingkat kemampuannya apakah mereka mengambil tes yang satu atau yang lainnya.

Penyetaraan bersifat simetri, artinya penyetaraan tes tidak bergantung pada tes mana yang digunakan sebagai tes rujukan (Hambleton, Swaminathan, & Rogers, 1991: 125). Fungsi yang digunakan untuk mentransfer skor pada tes 1 ke skala tes 2, sama dengan fungsi yang digunakan untuk mentransfer skor pada tes 2 ke skala tes 1. Penyetaraan bersifat grup invarian, artinya bahwa prosedur penyetaraan bebas sampel. Relasi pada penyetaraan harus sama tanpa memperhatikan grup peserta yang digunakan untuk melakukan penyetaraan (Petersen, Kolen, & Hoover, 1989).

Di antara sifat-sifat penyetaraan tersebut, sifat invarianlah yang sukar dipenuhi bila menggunakan metode penyetaraan klasik, karena sifat ini bertolak belakang dengan sifat teori tes klasik. Secara teori, metode penyetaraan dalam teori respon butir dapat mengatasi masalah ini. Jika penyetaraan berhasil, maka ketiga sifat ini terpenuhi, paling tidak mendekati terpenuhi.

4. Desain Penyetaraan

Terdapat tiga desain dasar yang dapat digunakan untuk pengumpulan data dalam melakukan penyetaraan tes. Desain yang dimaksud adalah desain grup tunggal, desain grup ekuivalen, dan desain tes jangkar (Hambleton, & Swaminathan, 1985: 198). Pemilihan desain yang digunakan dalam penyetaraan sangat tergantung pada situasi pelaksanaannya. Pada penelitian ini, dalam pengumpulan datanya menggunakan desai tes jangkar.

Bila desain tes-jangkar hendak digunakan, satu hal yang perlu mendapat perhatian yaitu butir-butir tes-jangkar harus mewakili tes total baik mengenai karakteristik, isi, dan statistiknya. Butir-butir tes-jangkar merupakan versi mini dari butir-butir tes total dan sebaiknya penempatannya berada pada posisi yang sama untuk kedua tes ketika tes tersebut diujikan. Dengan demikian tes-jangkar harus mewakili tes total.

Pada dasarnya hasil estimasi parameter butir-butir tes-jangkar, digunakan untuk menyetarakan estimasi parameter butir dan kemampuan dari kedua tes kedalam skala umum. Banyaknya butir jangkar yang digunakan perlu dipertimbangkan baik mengenai isi maupun bagian statistiknya. Mengenai isi, jelas butir-butir jangkar harus mempunyai spesifikasi yang sama, proporsional, dan mewakili tes total. Menurut Budescu (Kolen, & Brenan, 1995: 248) semakin banyak butir jangkar yang digunakan, semakin kecil kesalahan baku penyetaraannya. Ada yang mengatakan bahwa banyaknya butir jangkar boleh sedikit asal datanya unidimensi.

Kondisi di lapangan, tes pendidikan ada kecenderungan datanya heterogen, maka untuk memperoleh hasil penyetaraan yang memadai diperlukan banyaknya butir jangkar yang lebih besar. Berdasarkan pengalaman dapat ditetapkan bahwa banyaknya butir jangkar minimal 20 % dari panjang tes total yang memuat 40 butir atau lebih, kecuali tes yang sangat panjang, panjang tes-jangkar 30 sudah cukup (Kolen, & Brennan, 1995: 248).

5. Metode Penyetaraan

Dalam teori respon butir, jika model respon butir cocok dengan suatu set data, maka sebarang transformasi linear dari skala pengukuran juga cocok untuk data tersebut. Hal ini berarti, bahwa hubungan antara skala pengukuran dari dua tes adalah linear. Dengan demikian, jika skala tes 1 disetarakan dengan skala tes 2 untuk model 3-PL, maka hubungan parameter butir dan kemampuan peserta untuk dua skala tersebut dapat dinyatakan sebagai berikut (Kolen, & Brennan,1995: 165).

$$\theta_{2i} = \alpha\theta_{1i} + \beta, \qquad (2.1)$$

$$a_{2j} = \frac{a_{1j}}{\alpha} \qquad (2.2)$$

$$b_{2j} = \alpha\, b_{1j} + \beta \qquad (2.3)$$

dan

$$c_{2j} = c_{1j} \qquad (2.4)$$

dengan

a_{1j}, b_{1j}, dan c_{1j} adalah parameter butir untuk butir j pada skala tes 1,

a_{2j}, b_{2j}, dan c_{2j} adalah parameter butir untuk butir j pada skala tes 2,

θ_{1i} dan θ_{2i} berturut-turut merupakan kemampuan peserta i pada skala tes 1 dan tes 2,

α dan β adalah konstanta penyetaraan.

Untuk parameter tebakan c tidak ditransformasikan karena nilainya tidak bergantung pada metrik θ atau c bebas dari transformasi skala (Kolen &

Brennan, 1995: 163; Kaskowitz, & De Ayala, 2001: 40). Selanjutnya konstanta penyetaraan α dan β dapat dihitung dengan berbagai metode yang disebut metode penyetaraan. Dalam teori respon butir, metode penyetaraan yang diterapkan berasumsi bahwa kedua tes yang disetarakan unidimensi.

Terdapat berbagai metode yang dapat diterapkan untuk menyetarakan tes, antara lain metode Rerata & Sigma, Rerata & Sigma Tegar, Rerata & Rerata, dan Kurva Karakteristik. Diantara keempat metode tersebut, terdapat dua metode yang perlu mendapat perhatian yaitu metode Rerata & Sigma (RS) dan Rerata & Rerata (RR). Kedua metode tersebut formulanya sederhana dan penerapannya mudah. Metode yang populer adalah metode kurva karakteristik dari Stocking & Lord, karena metode tersebut yang dipandang terbaik di kelasnya. Berikut dibahas beberapa metode penyetaraan antara lain metode RS, RR, dan KK.

a. Metode Rerata & Sigma (RS)

Pada metode RS, untuk menentukan konstanta penyetaraan α dan β melibatkan rerata dan simpangan baku dari parameter tingkat kesulitan yang dapat dijelaskan sebagai berikut (Hambleton, Swaminathan, & Rogers, 1991: 129). Misal format tes 1 disetarakan dengan format tes 2. Karena hubungan parameter tingkat kesulitan berhubungan linear,

$$b_2 = \alpha b_1 + \beta ,$$

maka didapat

$$\overline{b_2} = \alpha \overline{b_1} + \beta ,$$

dan

$$S_2 = \alpha S_1$$

Jadi

$$\alpha = \frac{S_2}{S_1} , \qquad (2.5)$$

dan

$$\beta = \overline{b_2} \_ \alpha \overline{b_1}, \qquad (2.6)$$

dengan

$\overline{b_1}$ dan $\overline{b_2}$ adalah berturut-turut rerata tingkat kesulitan butir tes 1 dan tes 2,

S_1 dan S_2 adalah berturut-turut simpangan baku tingkat kesulitan butir tes 1 dan tes2

α dan β konstanta penyetaraan.

Setelah α dan β dihitung dengan menggunakan persamaan (2.5) dan (2.6), hasil estimasi parameter butir dan kemampuan dari tes 1 ditempatkan pada skala yang sama dengan tes 2 menggunakan hubungan sebagai berikut (Hambleton, Swaminathan, & Rogers, 1991: 129).

$$b_2^* = \alpha b_1 + \beta, \qquad (2.7)$$

$$a_2^* = \frac{a_1}{\alpha}, \qquad (2.8)$$

$$\theta_2^* = \alpha\theta_1 + \beta, \qquad (2.9)$$

dengan

b_2^* : nilai tingkat kesulitan butir-butir dalam tes1 ditempatkan pada skala tes 2,

a_2^* : nilai daya beda butir-butir dalam tes 1 ditempatkan pada skala tes 2.

θ_2^*: nilai kemampuan para peserta tes 1 ditempatkan pada skala tes 2.

Untuk butir-butir jangkar hasil estimasinya merupakan rata-rata hasil estimasi dari kedua tes, karena mereka tidak identik sebagai akibat dari kesalahan estimasi. Persamaan (2.7), (2,8), dan (2.9) disebut rumus konversi penyetaraan untuk model dikotomos. Kelebihan metode RS adalah formulanya sederhana, komputasinya dapat dilakukan secara manual atau bantuan program komputer sederhana. Kekurangan metode RR, dalam estimasi konstanta penyetaraan tidak melibatkan semua parameter butir.

b. Metode Rerata & Rerata (RR)

Pada metode RR, untuk menentukan konstanta penyetaraan melibatkan dua parameter butir, yaitu parameter daya beda dan tingkat kesulitan. Menurut Loyd & Hoover (Ogasawara, 2001a: 53) konstanta penyetaraan α dan β, dapat dihitung dengan menggunakan menggunakan rerata dari parameter butir yang terlibat yakni daya beda dan tingkat kesulitan butir. Misalkan penyetaraan dilakukan dari tes 1 ke tes 2, dengan desain tes-jangkar. Hubungan parameter tingkat kesulitan dan daya beda adalah sebagai berikut (Ogasawara, 2001a: 58).

$$b_2 = \alpha b_1 + \beta \,,\; a_2 = \frac{a_1}{\alpha}$$

Selanjutnya didapat

$$\overline{b_2} = \alpha \overline{b_1} + \beta \,,$$

dan

$$\overline{a_2} = \frac{\overline{a_1}}{\alpha}$$

Jadi

$$\alpha = \frac{\overline{a_1}}{\overline{a_2}} \qquad (2.10)$$

dan

$$\beta = \overline{b_2} \_ \alpha \overline{b_1} \,, \qquad (2.11)$$

dengan

$\overline{b_1}$, $\overline{b_2}$ adalah berturut-turut rerata tingkat kesulitan butir jangkar tes 1 dan tes 2,

$\overline{a_1}$, $\overline{a_2}$ adalah berturut-turut rerata daya beda butir jangkar tes1 dan tes 2,

α dan β konstanta penyetaraan.

Setelah α dan β didapat dari persamaan (2.10) dan (2.11), hasil estimasi parameter butir dan kemampuan dari tes 1 dikonversikan dengan hasil estimasi

parameter butir dan kemampuan dari tes 2, menggunakan rumus konversi pada persamaan (2.7)-(2.9). Kelebihan metode RR ini adalah formulanya sederhana, mudah diterapkan, dan komputasinya dapat dilakukan dengan cara manual atau bantuan program komputer sederhana. Kekurangan metode RR, dalam estimasi konstanta penyetaraan tidak melibatkan semua parameter butir.

e. Metode Kurva Karakteristik Haebara (HA)

Dalam menentukan konstanta penyetaraan pada metode RS maupun RR tidak melibatkan semua parameter butir secara simultan. Terdapat suatu metode penyetaraan yang melibatkan semua parameter butir secara simultan seperti yang diusulkan oleh Haebara dan dikembangkan oleh Stocking & Lord yaitu metode kurva karakteristik. Metode ini mempunyai 2 variasi formula dalam penerapannya.

Variasi yang pertama, konstanta penyetaraan dihitung dengan cara menentukan selisih antara setiap kurva karakteristik butir pada dua skala yang sudah disetarakan, dikuadratkan kemudian dijumlahkan (metode Haebara). Variasi yang kedua yaitu konstanta penyetraan dihitung dengan cara menentukan selisih antara kurva karakteristik tes dari dua skala yang sudah disetarakan lalu dikuadratkan (metode Stocking & Lord). Selanjutnya dengan menggunakan criteria tertentu, konstanta penyetaraan α dan β didapat dengan meminimumkan fungsi tertentu yang memuat variabel α dan β. Berikut ini. masing-masing metode disajikan.

Metode Haebara adalah metode kurva karakteristik yang formula komputasinya menggunakan variasi yang pertama, yang dapat dijelaskan sebagai berikut (Kolen, & Brennan, 1995: 170). Jumlah kuadrat dari selisih antara setiap kurva karakteristik butir dari dua skala yang sudah disetarakan adalah:

$$H(\theta_i) = \sum_{j=1}^{n} \left(T_{ij} - T_{ij}^{*}\right)^2 , \qquad (2.12)$$

dengan

$T_{ij} = P_j(\theta_i)$,

$T_{ij}^* = P_j^*(\theta_i)$,

dan

n : panjang tes-jangkar,

$P_j(\theta_i)$: Probabilitas peseta berkemampuan θ_I menjawab benar butir j,

$P_j^*(\theta_i)$: Probabilitas hasil transformasinya.

serta transformasi pada tes-jangkar,

$b_j^* = \alpha b_j + \beta$, $a_j^* = \frac{a_j}{\alpha}$, dan $c_j^* = c_j$.

Didefinisikan fungsi

$$F = \frac{1}{N}\sum_{i=1}^{N} H(\theta_i) = \frac{1}{N}\sum_{i=1}^{N}\ \sum_{i=1}^{n}\left(T_{ij} - T_{ij}^*\right)^2 \qquad (2.13)$$

dengan N : sebarang himpunan titik-titik sepanjang skala θ (Tate, 2000: 332). Fungsi F pada persamaan (2.13) merupakan fungsi dari α dan β, karena $(T_{ij} - T_{ij}^*)$ merupakan fungsi dari α dan β . Selanjutnya konstanta penyetaraan α dan β dipilih sedemikian rupa sehingga fungsi F minimum.

d. Metode Kurva Karakteristik Stocking & Lord (SL)

Terdapat suatu metode penyetaraan yang melibatkan semua parameter butir secara simultan seperti yang diusulkan oleh Haebara dan dikembangkan oleh Stocking & Lord yaitu metode kurva karakteristik..

Formula komputasinya dapat diturunkan sebagai berikut (Kolen, & Brennan, 1995: 170). Kuadrat dari selisih antara kurva karakterik tes dua skala yang disetarakan dinyatakan oleh:

$$SL(\theta_i) = (T_i - T_i^*)^2, \qquad (2.14)$$

dengan

$$T_i = \sum_{j=1}^{n} P_j(\theta_i)$$

$$T_i^* = \sum_{i=1}^{n} P_j^*(\theta_i)$$

dan

n : panjang tes-jangkar,

$P_j(\theta_i)$: Probabilitas peserta berkemampuan θi menjawab benar butir j,

$P_j^*(\theta_i)$: Probabilitas hasil transformasinya.

T_i : Skor tulen peserta berkemampuan θ_i pada tes dasar,

T_i^* : Skor tulen hasil tranformasinya.

Dengan transformasi pada tes-jangkar,

$b_j^* = \alpha b_j + \beta$, $a_j^* = \frac{a_j}{\alpha}$, dan $c_j^* = c_j$.

Di definisikan fungsi

$$F = \frac{1}{N}\sum_{i=1}^{N}(T_i - T_i^*)^2 \qquad (2.15)$$

Fungsi F pada persamaan (2.15) merupakan fungsi dari α dan β, karena $(T_i - T_i^*)$ merupakan fungsi dari α dan β. Selanjutnya konstanta penyetaraan α dan β dipilih sehingga fungsi F minimum. Fungsi F pada persamaan (2.13) dan (2.15), mencapai minimum bila

$$\frac{\partial F}{\partial \alpha} = \frac{\partial F}{\partial \beta} = 0 . \qquad (2.16)$$

Persamaan (2.16) non linear dan mempunyai solusi numerik, dapat diselesaikan dengan menggunakan prosedur numerik salah satu diantaranya adalah metode numerik Newton Raphson seperti yang direkomendasikan oleh Hambleton & Swaminathan (1985: 210)

6. Penelitian yang Relevan

Beberapa penelitian mengenai penyetaraan tes telah dilakukan, khususnya mengenai akurasi metode penyetaraan, baik model dikotomos, politomos, dan gabungannya. Hasil-hasil penelitian tentang penggunaan metode penyetaraan masih beragam, belum konsisten. Hal ini disebabkan tidak hanya semata-mata karena penerapan metode tersebut, tetapi mungkin akibat dari faktor-faktor yang ditinjau atau kondisi-kondisi yang dikembangkan oleh peneliti.

Ogasawara (2001a: 63) telah melakukan penelitian tentang perbandingan metode penyetaraan untuk model dikotomos (2-PL, 3-PL) dengan menentukan kesalahan baku asimtotik dari estimasi konstanta penyetaraan melalui data simulasi dan real. Metode yang dibandingkan adalah metode metode kurva karakteristik dan metode momen, hasilnya menunjukkan bahwa metode kurva karakteristik lebih akurat dari pada dari pada metode momen, khususnya model 3-PL. Secara umum kesalahan baku estimasi parameter butir untuk model 3-PL lebih besar dari kesalahan baku estimasi parameter butir unruk model 2-PL, tetapi kesalahan baku asimtotik estimasi konstanta penyetaraan dengan metode kurva karakteristik untuk model 3-PL dan 2-PL hampir sama.

Penelitian lain, membandingkan tiga metode yaitu metode kurva karakteristik, metode rerata sigma, dan metode kuadrat terkecil. Hasilnya menunjukkan bahwa metode kurva karakteristik mempunyai kesalahaan baku terkecil, disusul metode kuadrat terkecil dan metode rerata sigma. Satu keuntungan dari metode kuadrat terkecil adalah kesalahaan baku asimtotik dari estimasi konstanta penyetaraan mudah diturunkan dibandingkan dengan penurunan kesalahan baku dari estimasi konstanta penyetaraan dengan metode kurva karakteristik (Ogasawara, 2001b: 382).

Penelitian yang lain lagi, membandingkan tiga metode yaitu metode rerata sigma, metode kurva karakteristik, dan metode kalibrasi simultan

melalui penyetaraan skor tulen untuk model dikotomos 3-PL. Hasilnya menunjukkan bahwa kesalahan baku metode kurva karakteristik hampir sama dengan keslahan baku metode kalibrasi simultan, kesalahan baku metode rerata sigma lebih besar dari kesalahan baku dari dua metode lainnya (Ogasawara, 2001: 44). Berdasarkan penelitian yang dilakukan oleh Ogasawara, dapat disimpulkan bahwa metode kurva karakteristik lebih baik dari pada metode momen, metode kurva karakteristik hampir sama dengan metode kalibrasi simultan.Hasil penelitian Ogasawara ini tidak sejalan dengan hasil penelitian lain, misalnya penelitian yang dilakukan oleh Hanson & Beguin.

Hanson & Beguin juga melakukan penelitian mengenai perbandingan metode penyetaraan, yaitu membandingkan metode kalibrasi terpisah dan kalibrasi simultan, dan berbagai faktor yang mempengaruhi kestabilan hasil penyetaraannya antara lain: program estimasi yang digunakan, ukuran sampel, panjang tes-jangkar, dan ekuivalensi grup. Untuk metode kalibrasi terpisah membandingkan dua metode kurva karakteristik (Haebara dan Stocking & Lord) dan metode momen (RS dan RR). Hasilnya menunjukkan bahwa : 1) metode kurva karakteristik lebih baik dari metode momen. Prosedur Haebara dan Stocking & Lord menghasilan MSE (*mean squared error*) relatif sama, dan secara konsisten tidak ada metode yang lebih baik diantara keduanya: 2) ukuran sampel dan panjang tes-jangkar mempengaruhi kestabilan hasil penyetaraan, semakin besar ukuran sampel dan semakin banyak butir tes-jangkar semakin kecil MSEnya; 3) pada kondisi perbedaan rata-rata distribusi kemampuan antar grup 0, didapat kesalahan metode kalibrasi simultan lebih kecil dari kesalahan metode kalibrasi terpisah, baik menggunakan program BILOG-MG maupun MULTILOG; 4) pada kondisi perbedaan rata-rata distribusi kemampuan antar grup 1, didapat kesalahan metode kalibrasi simultan lebih kecil dari kesalahan metode kalibrasi terpisah dengan program BILOG-MG, dan kesalahan metode kalibrasi simultan lebih besar dari

kesalahan metode kalibrasi terpisah dengan program MULTILOG (Hanson, & Beguin, 2002: 12). Hasil penelitian yang dilakukan oleh Hanson & Beguin ini tidak sesuai dengan hasil-hasil penelitian sebelumnya. Kenyataan ini menunjukkan bahwa ditinjau dari faktor tertentu, suatu metode penyetaraan lebih stabil dari pada metode penyetaraan yang lain tetapi belum tentu demikian, jika ditinjau dari faktor yang lain. Dengan kata lain, kestabilan suatu metode penyetaraan dipengaruhi oleh berbagai faktor. Khusus untuk model polytomos (GRM), hasil penyetaraan tes dipengaruhi oleh ukuran sampel,banyaknya butir ancor, dan distribusi kemampuan testee (Swediati, 1997).

Ketidakstabilan suatu metode penyetaraan ini juga diperkuat oleh hasil penelitian yang dilakukan oleh Miyatun & Mardapi (2000: 16), yaitu dengan menentukan kesalahan bakunya, mengatakan bahwa metode momen lebih baik dari metode kurva karakteristik, kontradiksi dengan rekomendasi dari Kolen, & Brennan (1995: 174), bahwa metode kurva karakteristik lebih unggul dari metode momen. Hal ini menunjukkan bahwa hasil-hasil penelitian mengenai perbandingan metode penyetaraan tes, belum cukup untuk merekomendasikan bahwa suatu metode penyetaraan lebih unggul dari metode penyetaraan yang lain. Dengan demikian penelitian mengenai perbandingan metode penyetaraan masih perlu dilakukan.

7. Kerangka Pikir

Jelas bahwa nilai konstanta penyetaraan dihasilkan dari penggunaan metode penyetaraan tertentu. Masing-masing metode penyetaraan mempunyai kelebihan dan kekurangan yang satu sama lain berbeda. Oleh karena itu kestabilan hasil penyetaraan dipengaruhi oleh metode penyetaraan yang digunakan.

Formula perhitungan konstanta penyataraan untuk masing-masing metode,melibatkan hasil estimasi parameter butir. Kestabilan hasil estimasi

butir dipengaruhi oleh ukuran sampel. Dengan demikian kestabilan hasil penyetaraan juga dipengaruhi oleh ukuran sampel.

C. METODE PENELITIAN

Tujuan penelitian ini adalah untuk mengetahui metode penyetaraan tes yang paling stabil dalam penyetaraan tes berbentuk uraian. Penelitian ini menggunakaan data simulasi yang dibangkitkan berdasarkan data real. Untuk melakukan penelitian ini dapat dilakukan langkah-langkah sebagai berikut.

1. Pengumpulan data

Data dalam penelitian mengenai penyetaraan tes berbentuk uraian ini , adalah data empirik yang sudah terolah berupa hasil estimasi parameter butir dua perangkat tes berbentuk uraian terdiri dari 5 butir. Diantara 5 butir tes tersebut terdapat 3 butir yang sama sebagai butir-butir tes jangkar. Berdasarkan data empirik ini dibangkitkan data simulasi untuk keperluan replikasi.

2. Pembangkitan Data

Data respon siswa terhadap tes dibangkitkan berdasarkan data empirik dengan menggunakan program PASCAL. Data dibangkitkan berdasarkan model GRM dengan kondisi yang diperlukan dalam simulasi penelitian. Faktor yang akan diselidiki, yaitu ukuran sampel.Ukuran sampel menggunakan dua tingkat (500, 1000). Banyaknya butir tes-jangkar 40% dari banyaknya butir tes. Masing-masing kondisi dilakukan 25 replikasi.

3. Analisis Data

Dengan menggunakan program PARSCALE masing-masing set data untuk format tes 1 dan tes 2 dikalibrasi secara terpisah. Dari hasil kalibrasi ini, parameter butir dari format tes 1 disetarakan dengan parameter butir dari format tes 2 dengan menggunakan empat metode yaitu metode RS, RR, HA, dan SL. Untuk proses penyetaraan, komputasinya menggunakan program

STUIRT. Selanjutnya untuk mengetahui metode penyetaraan tes yang paling stabil, digunakan criteria tertentu.

4. Kriteria Evaluasi

Pada penentuan kestabilan hasil penyetaraan dari ketiga metode, dapat dilakukan dengan menghitung *root mean square differences* (RMSD) untuk kemampuan, yaitu RMSD antara parameter kemampuan hasil estimasi dan parameter kemampuan bangkitan. RMSD untuk kemampuan didefinisikan sebagai berikut (Kim & Cohen, 2002: 31).

$$RMSDuntuk\theta = \sqrt{\tfrac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2} = \sqrt{\tfrac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \overline{\theta})^2} + \sqrt{(\overline{\theta} - \theta_i)^2} \qquad (3.1)$$

dengan

N: banyaknya peserta,

$\hat{\theta}_i$: parameter kemampuan peserta ke i hasil estimasi pada grup sasaran,

θ_i: parameter kemampuan peserta ke I hasil bangkitan pada grup sasaran,

$\overline{\theta}$: rata-rata parameter kemampuan hasil estimasi pada grup sasaran.

RMSD untuk kemampuan, dapat dinyatakan dalam bentuk dekomposisi seperti pada ruas kanan persamaan (3.1), masing-masing sebagai simpangan baku dan bias.

Selanjutnya dengan menggunakan nilai rata-rata RMSD pada persamaan (3.1) menurut banyaknya replikasi, dapat ditentukan kestabilan dari keempat metode penyetaraan ditinjau dari ukuran sampel yang digunakan. Kriterianya adalah nilai rata-rata RMSD yang lebih kecil, menunjukkan bahwa metode penyetaraan lebih stabil.

D. HASIL DAN PEMBAHASAN

1.Hasil

Tabel berikut menyajikan hasil perhitungan rata-rata RMSD untuk parameter kemampuan dengan ukuran sampel 500 dan 1000, banyaknya butir tes jangkar 40 % untuk masing-masing metode penyetaraan.

Tabel 1.

Rata-rata RMSD untuk kemampuan dengan ukuran sampel 500 dan banyaknya butir tes jangkar 40 %.

No.	Metode Penyetaraan	Rata-rata RMSD Kemampuan	Standar Deviasi (SD)
1.	Rerata & Sigma (RS)	0.290970	0.035765
2.	Rerata & Rerata (RR)	0.318291	0.006212
3.	Haebara (HA)	0.213117	0.020829
4	Stocking & Lord (SL)	0.20661	0.020123

Tabel 2.

Rata-rata RMSD untuk kemampuan dengan ukuran sampel 1000 dan banyaknya butir tes jangkar 40 %.

No.	Metode Penyetaraan	Rata-rata RMSD Kemampuan	Standar Deviasi (SD)
1.	Rerata & Sigma (RS)	0.359746	0.056464
2.	Rerata & Rerata (RR)	0.235585	0.085884
3.	Haebara (HA)	0.321460	0.010241
4.	Stocking & Lord (SL)	0.293547	0.010122

2. Pembahasan

Berdasarkan Tabel 1., nilai rata- rata RMSD untuk kemampuan dengan ukuran sampel 500, berturut-turut mulai dari yang terkecil berasal dari metode SL, HA, RR dan RS. Demikian juga untuk ukuran sampel 1000 pada Tabel 2., menunjukkan bahwa nilai rata-rata RMSD untuk kemampuan berturut-turut mulai dari yang terkecil berasal dari metode RR, SL, HA, dan RS. Jadi baik ukuran ukuran sampel 500 maupun 1000 urutan kestabilan keempat metode tidak sama. Hal ini menunjukkan bahwa dengan ukuran sampel 500 dan 1000 , kekonsistenan kestabilan keempat metode tersebut belum tampak.

Kemudian jika ditinjau dari masing-masing metode dan ukuran sampel yang digunakan, dapat dikatakan bahwa, nilai RMSD untuk keempat metode dengan ukuran sampel 500 justru lebih kecil dari nilai RMSD untuk ukuran sampel 1000. Hal ini berarti metode lebih stabil untuk ukuran sampel 500 dari pada untuk ukuran sampel 1000.

Secara teori maupun berdasarkan hasil penelitian, semakin besar ukuran sampel, semakin stabil hasil estimasi parameter butir dan kemampuan. Hasil estimasi parameter butir dan kemampuan terlibat langsung dalam perhitungan konstanta penyetaraan. Akibatnya kestabilan metode penyetaraan dipengaruhi oleh ukuran sampel. Semakin besar ukuran sampel, semakin stabil metode penyetaraan tes yang digunakan. Hasil penelitian ini, tidak mendukung pernyataan tersebut, hal ini mungkin disebabkan karena perbedaan ukuran sampel yang digunakan antara ukuran sampel besar (1000) dan kecil (500) masih terlalu kecil, sehingga hasil estimasinya boleh dikatakan kurang stabil.

Terjadi ketidak konsistenan kestabilan metode penyetaraan untuk ukuran sampel 500 dan 1000. Untuk memastikan pengaruh ukuran sampel pada kestabilan metode penyataraan tes dapat dilakukan penelitian serupa dengan ukuran sampel yang lebih besar. Demikian juga pengaruh faktor-faktor yang lain, misalnya panjang tes, dan panjang tes jangjar

D. KESIMPULAN DAN SARAN

1. Kesimpulan

Kestabilan metode penyetaraan tes berbentuk uraian dengan menggunakan ukuran sampel 500 dan 1000 belum tampak. Ukuran sampel 500 dan 1000 pada penyetaraan tes berbentuk uraian pengaruhnya belum konsisten. Dengan kata lain untuk mendapatkan hasil penyetaraan tes yang stabil diperlukan ukuran sampel yang lebih besar.

2. Saran

Lakukan penelitian ulang dengan menggunakan ukuran sampel yang berbeda dan penyelidikan bisa dikembangkan untuk faktor yang lain misalnya banyaknya bitir ancor atau banyaknya kategori butir.

E. DAFTAR PUSTAKA

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Indiana: Prentice Hall.

Gronlund, N. E. (1976). Measurement and evaluation in teaching. New York: Macmillan Publishing Co.

Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of IRT in test contruction. *Applied Measurement in Education*, 4, 297-312.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Boston,MA: Kluwer Inc.

Kaskowitz, G. S., & De Ayala, R. J. (2001). The Effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25, 39-53.*

Kim, S-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the grade response model. *Applied Psychological Measurement, 26, 25-41.*

Kim, S., & Kolen, M. J. (2004). *STUIRT a computer program for scale tranformation under unidimensional item response theory models. v.1.0*. Diambil pada tanggal 8 Agustus 2006, dari http:// www.uiowa.edu/casma.

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

Lee, G., Kolen, M .J., Frisbie, D. A, at al. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25, 357-372.*

Muraki, E., & Bock, R. D. (1993). *PARSCALE: IRT item analysis and test scoring for rating-scale data.* Chicago, IL: Scientific Software International.

Miyatun, E., & Mardapi, D. (2000). Komparasi metode penyetaraan tes menurut teori respon butir. *Jurnal Penelitian dan Evaluasi, 3, 1-11.*

Ogasawara, H. (2001). Item response theort true score equatings and their standard errors. *Jurnal of Educational and Behavioral Statistic, 26, 31-50.*

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

Ogasawara, H. (2001b). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement, 25, 373-383.*

Petersen, N. S., Kolen, M. J.,& Hoover, H. D. (1989). Scaling, norming, and equating. Dalam Robert L. Linn(Ed.). *Educational Measurement (3rd ed.).* New York, NY: Macmillan.

Suryabrata, S. (1998). *Pengembangan alat ukur psikologis*. Yogyakarta: Direktorat Jenderal Pendidikan Tinggi Departemen Pendidikan dan Kebudayaan.

Swediati, N. (1997). *Equating tests under the generalized partial credit model*. A Disertation, University of Massachusetts Amherst.

Umar, J., Haribowo, H., Hayat, B., et al. (1997). *Bahan penataran pengujian pendidikan*. Jakarta: Depdikbud, Pusat Penelitian dan Pengembangan Sistem Pengujian.

Yin, P., Brennan, R. L., & Kolen, M. J. (2004). Concordance between ACT and ITED score from different populations. *Applied Psychological Measurement*, 28, 274-289.