EQUATING THE COMBINED DICHOTOMOUS AND POLITOMOUS ITEM TEST MODEL IN AN ACHIEVEMENT TEST

Kartono

Abstract

This study aims to reveal the significance and qualification of 1) the factor levels of the number of anchor items, the number of polytomous item categories, sample size, and 2) the combination of interfactor level influencing the results of test equating for the 3PL/GPCM mixed model. This study is a simulation study using empiric data generated from a mixed test model consisting of 30 multiple choice items and 5 essay items. The generated data was formed using five categories of polytomous items, with the sample sizes of 1000, 2000, 3000 and 25 replications. The analyses covered item analysis, constant of equating, RMSD evaluation criteria, test of significance factor, test of sinificant level difference, and qualification of level combination. Findings are as the followings. 1) The number of anchor items contributed significantly to the test equating, with 40% at the first level and 20% at the second level. The number of polytomous item categories significantly influences the test equating results at two levels, with 5 categories at the first level and 4 or 3 categories at the second level. 3) Sample size significantly influenced the test equating results, with 3000 at the first level, 2000 at the second level, and 1000 at the third level. Transformation model significantly influenced the test equating results, with the HA or SL method at the first level, RR at the second level, and RS at the third level.

Key words: *test equating, dichotomous, polytomous, anchor, category*

PENYETARAAN TES MODEL CAMPURAN BUTIR DIKOTOMUS DAN POLITOMUS PADA TES PRESTASI BELAJAR

Kartono

Abstrak

Penelitian ini bertujuan untuk mengungkapkan signifikansi dan kualifikasi: (1) level pada faktor banyaknya butir *anchor*, banyaknya kategori butir politomus, ukuran sampel, dan metode transformasi yang digunakan; dan (2) kombinasi level antar-faktor yang mempengaruhi hasil penyetaraan tes model campuran 3PL/GPCM. Penelitian ini merupakan penelitian simulasi. Data dibangkitkan berdasarkan data empirik dari tes bentuk campuran yang terdiri atas 30 butir pilihan ganda dan 5 butir uraian. Data bangkitan adalah butir politomus 5 kategori, ukuran sampel 1000, 2000, dan 3000, dan 25 replikasi. Analisis meliputi: analisis butir, konstanta penyetaraan, kriteria evaluasi dengan RMSD, uji signifikansi faktor, uji signifikansi perbedaan level, dan kualifikasi kombinasi level. Hasil penelitian adalah sebagai berikut. (1) Banyaknya butir *anchor* berpengaruh pada hasil penyetaraan tes, yaitu 40% pada level pertama dan 20% pada level kedua. (2) Banyaknya kategori butir politomus berpengaruh pada hasil penyetaraan tes pada dua level, yaitu 5 kategori pada level pertama, 4 atau 3 kategori pada level kedua. (3) Ukuran sampel berpengaruh pada hasil penyetaraan tes, yaitu 3000 pada level pertama, 2000 pada level kedua, dan 1000 pada level ketiga. (4) Metode transformasi berpengaruh pada hasil penyetaran tes, yaitu metode HA atau SL pada level pertama, RR pada level kedua, dan RS pada level ketiga.

Kata kunci: *penyetaraan tes, model dikotomus, politomus, anchor, kategori*

Pendahuluan

Peraturan Meteri Nomor 20 Tahun 2007 tentang Standar Penilaian Pendidikan mengamanatkan bahwa instrumen penilaian yang digunakan dalam Ujian Nasional (UN), di samping memenuhi persyaratan substansi, konstruksi, bahasa, dan memiliki bukti validitas empiris, juga menghasilkan skor yang dapat diperbandingkan antarsekolah, antardaerah, dan antartahun. Pada penyelenggaraan tes berskala besar seperti UN, demi keamanan, tes yang digunakan lebih dari satu paket. UN pada sekolah tingkat pendidikan dasar dan menengah, untuk setiap mata pelajaran yang diujikan, pernah menggunakan beberapa paket tes. Bahkan untuk keperluan UN tersebut, untuk setiap mata pelajaran yang diujikan pernah disiapkan tujuh paket tes, lima paket disediakan sebagai paket utama dan dua paket sebagai cadangan. Penggunaan beberapa paket dalam penyelenggaraan tes berskala besar juga diterapkan pada Ujian Masuk Perguruan Tinggi Negeri (UMPTN) di Indonesia (Mohandas, 2004: 1).

Pada situasi demikian, terdapat beberapa paket tes yang digunakan untuk mengukur variabel yang sama, namun skor hasil tes atau skor mentah yang dihasilkan tes tidak dapat langsung diperbandingkan walaupun tes-tes tersebut dibuat dengan kisi-kisi yang sama. Alasannya, skor mentah yang dihasilkan tes sering tidak memiliki titik nol yang sama atau ukuran yang sama. Selanjutnya, untuk menghasilkan skor tes yang dapat diperbandingkan dari beberapa paket tes, perlu disusun skor tes yang memiliki skala tunggal (*common scale*). Proses statistik yang digunakan untuk menghasilkan skala tunggal dari beberapa paket tes ini disebut penyetaraan tes (Kolen & Brennan, 1995: 2).

Konsep penyetaraan tersebut menjadi sangat urgen dengan munculnya kebijakan otonomi pendidikan yang mengakibatkan setiap daerah otonom mempunyai wewenang untuk menyelenggarakan pendidikan termasuk di dalamnya mengembangkan tes masing-masing. Kebijakan otonomi pendidikan ini masih memberi kewenangan kepada Pemerintah pusat untuk mengendalikan kualitas pendidikan nasional. Kewenangan daerah untuk melaksanakan pendidikan, memungkinkan diselenggarakannya tes prestasi belajar pada pendidikan dasar dan

menengah berskala besar misalnya Ulangan Akhir Semester (UAS) di tingkat daerah. Akibatnya terdapat beberapa paket tes yang berbeda (versi daerah) yang mengukur variabel yang sama. Pengendalian akan mudah dilakukan jika perangkat tes di tiap-tiap daerah disetarakan.

Di negara maju seperti Amerika Serikat, penyetaraan tes sangat diperlukan oleh lembaga-lembaga yang menangani tes khusus yaitu *test battery*, misalnya *Iowa Test of Basic Skills* (ITBS) dan *Iowa Test of Educational Development* (ITED). Tes tersebut digunakan untuk mengukur perkembangan kemampuan atau prestasi siswa pada jenjang dan kelas tertentu, terdiri dari beberapa tingkat yang terkait dengan umur dan kelas (Hieronymus, et al, 1980). Kelompok siswa yang mempunyai umur dan kelas yang sama, dalam perkembangannya, terdapat siswa yang cepat belajarnya dan ada yang lambat belajarnya.

Khusus untuk siswa yang cepat belajarnya, siswa tersebut cocok diberi tes tingkat lanjut, tetapi untuk anak yang lambat belajarnya, siswa tersebut tidak cocok diberi tes tingkat lanjut. Tidak perlu diadakan tes tingkat lanjut untuk siswa yang lambat belajar, karena tes yang terlalu sulit dapat menyebabkan siswa frustasi, cukup dilakukan penyesuaian skor pada tes tingkat di bawahnya untuk mengetahui kemampuan pada tes tingkat lanjut. Di Indonesia tes semacam itu pernah diadakan pada siswa kelas tiga sekolah dasar, tes kemampuan dasar membaca menulis berhitung yang sering disebut tes "calistung", tetapi belum melembaga.

Dengan memfasilitasi siswa-siswa dengan kemampuan yang berbeda, akan terdapat beberapa paket tes dengan tingkat yang berbeda, skala skor hasil tes berbeda. Perlu diadakan penyesuaian skor yang berasal dari hasil tes dengan tingkat yang berbeda pada skala yang sama, sehingga hasil tes dapat diperbandingkan. Dalam hal ini, untuk menyesuaikan skor-skor dari beberapa paket tes dengan tingkat yang berbeda agar skor tersebut dapat dibandingkan juga diperlukan penyetaraan tes.

Menurut Jahja Umar (1995: 9) bentuk tes yang digunakan dalam ujian tergantung dari jenis ujian. Jika jenis ujiannya adalah ulangan harian sebaiknya tes yang digunakan adalah tes bentuk uraian yang lebih detail. Jika jenis ujiannya adalah ujian akhir, tes yang digunakan adalah tes bentuk

pilihan ganda atau campuran antara pilihan ganda dan uraian sesuai dengan tuntutan kebutuhan yang ada.

Berdasarkan pengamatan di sekolah pada ujian akhir khususnya UN, biasa menggunakan tes berbentuk pilihan ganda, beberapa kali pernah menggunakan tes berbentuk campuran antara pilihan ganda dan uraian. Karena pertimbangan tertentu soal tes bentuk campuran ini tidak dipakai lagi dalam UN. Soal tes berbentuk campuran antara pilihan ganda dan uraian biasa digunakan pada UAS baik pada pendidikan dasar maupun menengah. Bentuk soal seperti ini memerlukan analisis dengan pendekatan teori tertentu untuk mengetahui kualitasnya.

Pada teori respons butir, soal berbentuk pilihan ganda termasuk butir dikotomus, sedangkan soal uraian termasuk butir politomus. Ada beberapa model respons butir yang dapat digunakan. Pada butir dikotomus, model yang sering digunakan adalah model logistik satu parameter, dua parameter, dan tiga parameter yang berturut-turut disingkat dengan 1-PL, 2-PL, dan 3-PL. Model logistik satu parameter sering disebut model Rasch. Di antara tiga model tersebut, model 3-PL yang sesuai dengan kondisi alami dalam pelaksanaan tes, yaitu adanya indeks daya beda butir yang berbeda dan tebakan butir dalam tes.

Pada butir politomus ada beberapa model penskoran. Model yang sering digunakan antara lain *The Graded Response Model* (GRM), *Nominal Response Model* (NRM), *Rating Scale Model* (RSM), *Partial Credit Model* (PCM), dan *Generalized Partial Credit Model* (GPCM). Khusus model GPCM ini, jika respons terdiri dari dua kategori dengan asumsi tidak ada tebakan, maka model tersebut menjadi model Rasch atau 2-PL. Jadi, model Rasch dan 2-PL merupakan kasus khusus dari model GPCM.

Format tes model campuran merupakan kombinasi antara model dikotomus dan politomus, dan penerapannya mengikuti asumsi yang sama dengan model pembentuknya yaitu unidimensi. Model 1-PL biasa dikombinasikan dengan PCM menjadi model campuaran 1-PL/PCM, sedangkan model-model 2-PL dan 3-PL dikombinasikan dengan GRM, dan GPCM, sehingga model campuran yang mungkin 2-PL/GRM, 2-PL/GPCM, 3-PL/GRM, dan 3-PL/GPCM. Menurut Fitzpatrick (Bastari, 1998: 2) analisis butir yang menggunakan model campuran 1-PL/PCM

hasilnya kurang memuaskan bila dibandingkan dengan model campuran 3-PL/GPCM. Penerapan model campuran 3-PL/GRM dan 3-PL/GPCM dalam analisis butir tes, hasilnya relatif sama (Bastari, 1998: 20), sehingga untuk analisis butir tes model campuran lebih baik menggunakan model campuran 3-PL/GRM atau 3-PL/GPCM.

Pada situasi pengujian tertentu, format tes model campuran antara pilihan ganda dan uraian biasa digunakan, sehingga memungkinkan pengembangan tes model tersebut termasuk penyetaraan tes. Menurut Sykes & Yen (2000: 221) penyetaraan tes model campuran antara pilihan ganda dan uraian lebih menguntungkan dari pada penyetaraan tes model tunggal secara terpisah. Beberapa keuntungan penyetaraan tes model campuran antara lain: (a) jika skor dari kedua tes berkorelasi positif, maka skor total dari tes model campuran lebih reliabel bila dibandingkan skor dari tes secara terpisah; (b) jika penyetaraan dilakukan pada kedua format tes terpisah maka hasil penyetaraannya tidak stabil; dan (c) penyetaraan tes model campuran memberikan hasil dalam skala tunggal.

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 yang digunakan. Dua hal tersebut, memiliki pengaruh yang sangat besar pada hasil penyetaraan. Oleh karena itu, untuk memaksimalkan kualitas hasil penyetaraan tes, perlu pemilihan desain dan metode penyetaraan yang tepat.

Pemilihan desain penyetaraan, tergantung pada format tes yang diujikan dan grup peserta yang diberi tes. Desain penyetaraan tes, diperlukan untuk merancang pengumpulan data dalam penyetaraan tes. Ada beberapa desain yang dapat digunakan untuk menyetarakan tes, antara lain, desain grup tunggal, desain grup ekuivalen, dan desain tes anchor. Di antara desain-desain tersebut, desain penyetaraan tes yang lebih menguntungkan dibandingkan dengan desain penyetaraan lainnya adalah desain tes anchor (Hambleton, Swainathan, & Rogers, 1991: 129). Pada desain ini dua grup peserta masing-masing diberi paket tes yang berbeda, tetapi mengukur variabel yang sama. Satu set butir anchor disisipkan ke

dalam masing-masing paket dan butir ini digunakan untuk mengaitkan beberapa paket tes dan menempatkan parameter butir dan kemampuan dari setiap paket pada satu skala.

Seperti halnya pada desain penyetaraan, pada metode penyetaraan juga terdapat berbagai metode yang dapat digunakan. Berdasarkan proses kalibrasinya, metode penyetaraan secara umum dapat dibedakan menjadi dua macam yaitu metode penyetaraan kalibrasi simultan dan terpisah. Pada metode kalibrasi simultan, nilai parameter butir dan kemampuan dari kedua tes sudah berada pada skala yang sama tanpa ada perhitungan konstanta penyetaraan, sedangkan pada metode kalibrasi terpisah, nilai parameter butir dan kemampuan dari kedua tes belum berada pada skala yang sama sehingga perlu disetarakan dengan menentukan konstanta penyetaraan.

Metode kalibrasi terpisah dibedakan menjadi metode momen dan metode fungsi respons tes. Untuk metode momen, ada empat metode yang bisa diterapkan yaitu metode regresi, metode rerata & sigma, metode rerata & sigma robus, dan metode rerata & rerata. Pada metode regresi, di samping hanya melibatkan parameter indeks kesukaran butir dari masing-masing tes, juga mempunyai keterbatasan bahwa sifat simetri dari penyetaraan tidak terpenuhi, sehingga metode regresi tidak disarankan untuk diterapkan. Metode rerata & sigma hanya melibatkan satu parameter yaitu parameter indeks kesukaran butir. Di samping rumusnya sederhana, semua sifat penyetaraan dipenuhi oleh metode ini. Metode rerata & sigma robus merupakan perluasaan dari metode rerata dan sigma yaitu dengan memberi pembobotan pada hasil estimasi parameter indeks kesukaran butir pada masing-masing tes. Seperti halnya metode rerata & sigma, metode rerata & rerata rumusnya juga sederhana dan sifat-sifat penyetaraan juga dipenuhi, dalam menentukan konstanta penyetaraan melibatkan dua parameter, yaitu parameter daya beda dan indeks kesukaran butir.

Pada metode fungsi respons tes, penentuan konstanta penyetaraan melibatkan semua parameter butirnya. Ada dua metode yang dapat diterapkan, yaitu metode kuadrat terkecil dan metode kurva karakteristik. Pada metode kuadrat terkecil, komputasi penentuan konstanta penyetaraan tanpa iterasi (Ogasawara, 2001b: 373). Pada metode kurva karakteristik,

komputasi konstanta penyetaraan memerlukan iterasi (Hambleton, Swainathan, & Rogers, 1991: 135; Kolen & Brennan, 1995: 170).

Terdapat dua metode yang termasuk metode kurva karakteristik, yaitu metode penyetaraan tes yang dikembangkan oleh Haebara yang selanjutnya disebut metode kurva karakteristik Haebara dan metode penyetaraan tes yang dikembangkan oleh Stocking & Lord yang selanjutnya disebut metode kurva karakteristik Stocking & Lord. Kedua metode penyetaraan tes tersebut sama-sama melibatkan semua parameter butirnya dalam penentuan konstanta penyetaraan, dan berbeda hanya pada variasi rumusnya. Beberapa metode penyetaraan yang telah disebutkan tadi, masing-masing mempunyai kelebihan dan kekurangan. Secara umum untuk model dikotomus, metode kurva karakteristik Stocking & Lord yang terbaik (Kolen & Brennan, 1995: 174).

Ketika desain yang hendak digunakan untuk keperluan pengumpulan data dalam penyetaraan tes adalah desain tes anchor, maka faktor banyaknya butir anchor sangat berperan dalam proses penentuan hasil penyetaraan tes. Parameter butir anchor inilah yang terlibat langsung dalam penentuan konstanta penyetaraan. Banyaknya butir anchor yang digunakan, bersesuaian dengan banyaknya nilai parameter butir yang terlibat dalam penentuan konstanta penyetaraan. Semakin banyak butir anchor yang digunakan, semakin besar frekuensi nilai parameter butir yang terlibat.

Nilai parameter butir yang terlibat dalam penentuan konstanta penyetaraan, diperoleh melalui estimasi. Hasil estimasi nilai parameter butir tesebut dipengaruhi beberapa faktor, antara lain ukuran sampel, sehingga faktor tersebut berperan juga dalam penentuan hasil penyetaraan tes. Dalam hal ini, faktor ukuran sampel juga perlu diungkap dalam kaitannya dengan hasil penyetaraan tes.

Selain banyaknya butir anchor dan ukuran sampel, faktor lain yang perlu diungkap juga adalah banyaknya kategori butir politomus. Banyaknya kategori butir akan menentukan banyaknya langkah dalam memberi respons pada butir dan untuk setiap langkah mempunyai indeks kesukaran kategori butir dan terlibat langsung pada penentuan konstanta penyetaraan. Banyaknya kategori butir, bersesuaian dengan banyaknya parameter butir.

Semakin besar kategori butir, semakin banyak parameter yang terlibat dalam penentuan konstanta penyetaraan.

Nilai konstanta penyetaraan ditentukan berdasarkan suatu metode penyetaraan. Penggunaan metode penyetaraan yang berbeda dapat memberikan nilai yang berbeda, karena masing-masing metode mempunyai kelebihan, kekurangan, dan rumus yang berbeda. Dengan demikian, penggunaan metode penyetaraan akan mempengaruhi kualitas hasil penyetaraan. Oleh karena itu, perlu pemilihan metode penyetaraan tes yang dapat memberikan hasil penyetaraan tes yang cermat.

Penelitian ini bertujuan untuk mengetahui signifikansi dan kualifikasi level pada faktor-faktor banyaknya butir anchor, banyaknya kategori butir politomus, ukuran sampel, dan metode penyetaraan dalam mempengaruhi hasil penyetaraan tes model campuran 3-PL/GPCM. Di samping itu, juga untuk mengetahui kualifikasi kombinasi level antar faktor (banyaknya butir anchor, banyaknya kategori butir politomus, ukuran sampel, dan metode penyetaraan) dalam mempengaruhi hasil penyetaraan tes model campuran 3-PL/GPCM.

Metode Penelitian

Penelitian dilakukan dengan menggunakaan data simulasi yang dibangkitkan berdasarkan data empirik. Sebelum penelitian simulasi dilakukan terlebih dahulu dilakukan penelitian empirik. Langkah-langkah dalam penelitian ini dapat dilakukan sebagai berikut.

1. Penelitian Empirik

Data dalam penelitian ini berupa respons siswa pada tes prestasi belajar yang biasa dilaksanakan pada akhir semester disebut Ulangan Akhir Semester (UAS) untuk mata pelajaran Matematika SMA kelas 2. UAS diselenggarakan dalam skala besar, dikoordinir oleh sekolah di bawah koordinasi seorang Kepala Sekolah sebagai koordinator Kelompok Kerja Kepala Sekolah (KKKS), diikuti oleh hampir semua siswa SMA Negeri dan Swasta di kota Semarang. Bentuk tes adalah campuran terdiri dari 30 butir

pilihan ganda dengan 5 pilihan jawaban dan 5 butir uraian dengan 5 kategori tiap butir.

Terdapat dua paket tes, yaitu paket tes 1 dan tes 2. Paket tes 1, untuk UAS yang diselenggarakan pada akhir semester genap tahun pelajaran 2004/2005, paket tes 2, untuk UAS yang diselenggarakan pada akhir semester genap tahun pelajaran 2003/2004. Tes disusun oleh tim guru yang tergabung dalam Musywarah Guru Mata Pelajaran (MGMP) tingkat wilayah berdasarkan Kurikulum 1994. Tes dibuat berdasarkan kisi-kisi tes dan analisis butir secara kualitatif, sehingga tes yang digunakan belum dikembangkan dengan baik. Kedua paket tes tersebut memuat butir-butir soal yang sama sebagai butir anchor. Banyaknya butir anchor untuk soal pilihan ganda 15 butir dan untuk soal uraian 3 butir. Kisi-kisi soal, soal ulangan akhir semester II tahun pelajaran 2004/2005, dan soal ulangan akhir semester II tahun pelajaran 2003/2004 berturut-turut terlampir pada Lampiran 2, Lampiran 3, dan Lampiran 4.

Pengumpulan data dilakukan melalui sekolah-sekolah, dengan mengopy respons siswa sebagai hasil tes dari masing-masing kelompok. Diperoleh dua data set, yaitu data set yang pertama adalah respons siswa pada UAS genap yang diselenggarakan pada tahun pelajaran 2004/2005 dengan 1224 responden dan data set yang kedua adalah respons siswa pada UAS genap yang diselenggarakan pada tahun pelajaran 2003/2004 dengan 1260 responden. Melalui analisis butir klasik, untuk memperoleh nilai parameter butir yang memenuhi syarat, responden yang tidak konsisten tidak diikutkan dalam analisis ini, sehingga untuk masing-masing kelompok diperoleh data berturut-turut dengan ukuran sampel 1192 dan 1178. Pada pengambilan data ini, memperhatikan penyebaran kemampuan siswa secara umum dalam arti tingkat kemampuan kelompok siswa terwakili.

Analisis butir dilakukan untuk menentukan nilai parameter butir dan kemampuan peserta dari kedua tes. Tujuannya disamping untuk keperluan pembangkitan data untuk simulasi, juga digunakan untuk memilih butir anchor yang akan digunakan.

Metode yang digunakan untuk analisis butir ini adalah metode Estimasi Bayes *(Bayesian Estimation)*, menggunakan model campuran 3-PL/GPCM. Metode tersebut dapat mengestimasi respons peserta berskor

sempurna atau skor nol (Hambleton & Swaminathan, 1985: 91). Komputasinya menggunakan program PARSCALE (Muraki & Bock, 1993).

2. Penelitian Simulasi

a. Variabel Penelitian

Sebagai variabel bebas (faktor) dalam penelitian ini adalah banyaknya butir anchor, banyaknya kategori butir, ukuran sampel, dan metode transformasi yang digunakan dalam penyetaraan tes. Sebagai variabel terikatnya adalah hasil penyetaraan tes model campuran 3-PL/GPCM yang diukur melalui nilai RMSD untuk kemampuan.

b. Diskripsi Faktor

Terdapat empat faktor yang ditinjau, yaitu banyaknya butir anchor, banyaknya kategori butir, ukuran sampel, dan metode transformasi yang digunakan dalam penyetaraan tes. Dalam simulasi, pemilihan level atau taraf untuk setiap faktor yang ditinjau diusahakan mendekati level yang didapat dalam situasi nyata. Banyaknya butir anchor, menggunakan dua level yaitu 20% dan 40% dari tes total, berturut-turut sebagai ukuran minimal yang disarankan dan sebagai ukuran sedang. Banyaknya kategori butir politomus, menggunakan tiga level yaitu tiap butir 3, 4, dan 5 kategori, sesuai dengan kategori yang digunakan untuk pembuatan butir soal politomus pada tes akhir semester untuk mata pelajaran Matematika SMA. Ukuran sampel, menggunakan tiga level yaitu 1000, 2000, dan 3000 berturut-turut sebagai ukuran sampel kecil, sedang, dan besar. Metode transformasi yang digunakan, menggunakan empat level, yaitu metode Rerata dan Rerata (RR), metode Rerata dan Sigma (RS), metode Haebora (HA), dan metode Stocking dan Lord (SL).

c. Disain Eksperimen

Penelitian ini menggunakan disain eksperimen faktorial (empat faktor), satu amatan untuk setiap sel (kombinasi level antarfaktor). Pada

desain eksperimen faktorial dengan satu amatan setiap sel berasumsi bahwa interaksi antarfaktor berderajat tinggi diabaikan dan digabungkan pada kesalahan eksperimen (Montgomery, 1984: 274). Interaksi antarfaktor berderajad 3 dan 4 pada disain empat faktor dianggap interaksi derajat tinggi. Oleh karena itu pada desain eksperimen dalam penelitian ini, interaksi antarfaktor berderajad 3 dan 4 diabaikan.

d. Pembangkitan Data

Data yang dibangkitkan adalah respons siswa untuk setiap grup pada setiap tes. Data dibangkitkan dengan kondisi panjang tes 35 yang terdiri 30 butir pilihan ganda dan 5 butir uraian dengan 5 kategori tiap butir, ukuran sampel 1000, 2000, dan 3000. Data dibangkitkan berdasarkan data empirik, melalui parameter butir menggunakan distribusi seragam dengan memperhatikan jangkaun masing-masing parameter yang direkomendasikan dan parameter kemampuan menggunakan distribusi normal, yang secara rinci disajikan Tabel 1 sebagai berikut.

Tabel 1. Distribusi Parameter Butir dan Kemampuan yang Digunakan untuk Pembangkitan Data

Butir	Parameter	Distribusi (Tes 1)	Distribusi (Tes 2)
Dikotomus	Daya beda	[0,37; 1,77]	[0,37; 1,64]
	Indeks kesukaran	[-0,72; 1,99]	[-0,68; 2,10]
	Tebakan	[0,00; 0,33]	[0,00; 0,31]
	Kemampuan	normal	normal
Politomus	Daya beda	[0,28; 0,42]	[0,37; 0,46]
	Indeks kesukaran	[-0,76; 1,33]	[-0,70; 1,49]
	Kemampuan	normal	normal

Data dibangkitkan berdasarkan model 3PL/GPCM menggunakan program WINGEN2. Program tersebut khusus untuk membangkitkan data respons butir model tunggal dikotomus dan politomus, maupun model campuran keduanya untuk beberapa model serta beberapa kondisi yang sesuai dengan kondisi nyata dalam praktik. Program WINGEN2 dapat digunakan untuk membangkitkan data respons butir dengan nilai parameter butir dan

kemampuan untuk berbagai distribusi yang sesuai dengan distribusi data nyata (Han & Hambleton, 2007: 5). Pada pembangkitan data ini, setiap kondisi dilakukan 25 replikasi.

Data dengan kondisi banyaknya kategori 3 dan 4, tidak perlu dibangkitkan, karena dengan data yang sudah ada (kondisi 5 kategori), dapat dilakukan analisis butir dengan kondisi 3 kategori atau 4 kategori. Perubahan banyaknya kategori dari 5 menjadi 4 dan 3, dilakukan langsung melalui modifikasi sintaks program yang digunakan untuk analisis butir.

Perubahan banyaknya kategori dari 5 menjadi 3, dilakukan pada sintaks program dengan memodifikasi skor pada 5 kategori dari (1,2,3,4,5) menjadi (1,1,2,3,3) pada tiga kategori, artinya skor 1 pada 5 kategori tetap 1 pada tiga kategori, skor 2 pada 5 kategori berubah menjadi skor 1 pada 3 kategori, skor 3 pada 5 kategori berubah menjadi skor 2 pada 3 kategori, dan skor 4 dan 5 pada 5 kategori berubah menjadi skor 3 pada 3 kategori. Demikian juga perubahan banyaknya kategori dari 5 menjadi 4, dilakukan dengan memodifikasi skor pada 5 kategori dari (1,2,3,4,5) menjadi (1,1,2,3,4) pada 4 kategori, artinya skor 1 pada 5 kategori tetap 1 pada 4 kategori, skor 2 pada 5 kategori berubah menjadi skor 1 pada 4 kategori, skor 3 pada 5 kategori berubah menjadi skor 2 pada 4 kategori, dan skor 4 pada 5 kategori berubah menjadi skor 3 pada 3 kategori, dan skor 5 pada 5 kategori berubah menjadi skor 4 pada 4 kategori.

e. Analisis Data

1) Analisis Butir

Pada tahap ini, analisis dilakukan untuk mendapatkan parameter butir (butir anchor) dan kemampuan dari setiap tes. Terdapat 9 kondisi data yang dianalisis untuk setiap tes. Kondisi data tersebut terbentuk dari faktor ukuran sampel 3 level dan banyaknya kategori 3 level.

Data untuk setiap kondisi pada setiap tes dianalisis secara terpisah dengan menggunakan bantuan program PARSCALE (Muraki & Bock, 1997). Dari analisis ini, diperoleh parameter butir dan kemampuan untuk setiap kondisi pada setiap tes. Selanjutnya hasil analisis ini, dipakai untuk menentukan konstanta penyetaraan.

2) Konstanta Penyetaraan

Seperti halnya pada analisis data empirik, konstanta penyetaraan diperlukan untuk menyetarakan parameter butir dan kemampuan dari kedua tes, sehingga parameter butir dan kemampuan dari kedua tes tersebut berada pada skala yang sama. Konstanta penyetaraan dihitung dengan menggunakan empat metode, yaitu metode RR, RS, HA, dan SL. Komputasinya menggunakan program STUIRT versi 1.0.

STUIRT adalah suatu program komputer untuk mengimplementasikan empat metode transformasi skala, yaitu metode RR, RS, HA, dan SL dalam menentukan konstanta penyetaraan tes. Program tersebut dapat digunakan untuk menghitung konstanta penyetaraan tes model tunggal dikotomus, politomus, dan campuran keduanya. Fasilitas model yang tersedia pada program untuk model dikotomus adalah model 1-PL, 2-PL, dan 3-PL, sedangkan untuk model politomus adalah GRM, GPCM, dan NRM (Kim & Kolen, 2004: 2).

3) Kriteria Evaluasi

Kualitas hasil penyetaraan antar level untuk setiap faktor yang ditinjau dapat ditentukan dengan membandingkan rata-rata nilai *root mean square difference* (RMSD) kemampuan dari masing-masing level untuk setiap faktor tersebut. Pada suatu suatu kondisi, RMSD kemampuan dapat ditentukan dengan menggunakan rumus sebagai berikut (Kim & Cohen, 2002: 31).

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

dengan

N : ukuran sampel,

$\hat{\theta}_i$: kemampuan peserta ke i setelah disetarakan,

θ_i : kemampuan peserta ke i sebelum disetarakan.

Kriteria uji yaitu rata-rata RMSD kemampuan yang nilainya lebih kecil menunjukkan bahwa kualitas hasil penyetaraan lebih baik. Komputasi RMSD pada (1) menggunakan program Microsoft Office EXCEL (Nana Suarna, 2003).

4) Uji Signifikansi Faktor

Selanjutnya untuk menguji signifikansi pengaruh faktor-faktor banyaknya butir anchor, banyaknya kategori butir, ukuran sampel, dan metode tranformasi yang digunakan, menggunakan ANAVA. Komputasinya menggunakan program SPSS 12 (Arif Pratisto, 2004).

5) Uji Signifikansi Perbedaan Level

Jika pada uji signifikansi faktor hasilnya signifikan, maka perlu dilakukan uji lanjut untuk mengetahui signifikansi perbedaan level pada masing-masing faktor, yaitu level-level yang mempunyai nilai rata-rata RMSD kemampuan yang perbedaannya signifikan. Uji ini menggunakan uji *least significance difference* (LSD) yang dinyatakan dengan rumus sebagai berikut (Montgomery, 1984: 65).

$$LSD = t_{\alpha/2,\, N-a} \sqrt{\frac{2RJK_S}{n}} \qquad (2)$$

dengan

t : statistik t,
α : taraf signifikansi (5%),
N : banyaknya data,
A : banyaknya level pada faktor yang ditinjau,
N : banyaknya data tiap level,
RJK_S : rata-rata jumlah kuadrat kesalahan.

Kriteria uji LSD, jika selisih nialai rata-rata RMSD kemampuan $|\bar{y}_i - \bar{y}_j| >$ LSD maka perbedaan nilai rata-rata RMSD kemampuan signifikan. Berdasarkan kriteria evaluasi, bahwa rata-rata RMSD kemampuan yang nilainya lebih kecil menunjukkan bahwa kualitas hasil

penyetaraan lebih baik. Kualifikasi level pada setiap faktor ditentukan berdasarkan kriteria evaluasi. Komputasi nilai LSD pada (2) menggunakan program SPSS 12 (Arif Pratisto, 2004).

6) Kualifikasi Kombinasi Level Antar-faktor

Berdasarkan hasil uji signifikansi perbedaan level untuk setiap faktor, didapat kualifikasi level. Jika kualifikasi level pada setiap faktor tersebut dikombinasikan, maka didapat kombinasi peringkat antarfaktor. Kualifikasi kombinasi level antarfaktor dapat ditentukan berdasarkan peringkat nilai rata-rata RMSD kemampuan yang bersangkutan.

Hasil Penelitian

Berdasarkan hasil penelitian dan pembahasan, dapat disimpulkan sebagai berikut. Faktor-faktor banyaknya butir anchor, banyaknya butir kategori butir, ukuran sampel, dan metode penyetaraan yang digunakan berpengaruh pada hasil penyetaraan tes model 3-PL/GPCM. Dua level yang ditinjau pada faktor banyaknya butir anchor dua-duanya merupakan level berbeda yang signifikan, dengan kualifikasi peringkat pertama 40% dan kedua 20%. Hasil penyetaraan tes yang melibatkan butir anchor 40% lebih baik dari hasil penyetaraan yang melibatkan butir anchor 20%.

Terdapat dua pasang level berbeda yang signifikan pada faktor banyaknya kategori butir politomus dari tiga level yang ditinjau, yaitu pasangan 5 K & 4 K dan pasangan 5 K & 3 K, dengan kualifikasi level peringkat pertama 5 K, dan kedua 4 K atau 3 K. Hasil penyetaraan tes yang melibatkan banyaknya kategori 5 K lebih baik dari hasil penyetaraan yang melibatkan banyaknya kategori 4 K dan 3 K. Hasil penyetaraan yang melibatkan banyaknya kategori 4 K dan 3 K kualitasnya sama.

Tiga level yang ditinjau pada faktor ukuran sampel tiga-tiganya merupakan level berbeda yang signifikan, dengan kualifikasi level peringkat pertama ukuran sample 3000, kedua ukuran sample 2000, dan ketiga ukuran sampel 1000. Hasil penyetaraan tes yang menggunakan ukuran sampel 3000 lebih baik dari hasil penyetaraan tes yang menggunakan ukuran sampel 2000. Hasil penyetaraan tes yang menggunakan ukuran

sampel 2000 lebih baik dari hasil penyetaraan tes yang menggunakan ukuran sampel 1000.

Pasangan level berbeda pada faktor metode transformasi yang digunakan dari empat level yang ditinjau semuanya signifikan kecuali pasangan level HA & SL, dengan kualifikasi level peringkat pertama metode SL atau metode HA, kedua metode RR, dan ketiga metode RS. Hasil penyetaraan tes yang menggunakan metode SL atau HA lebih baik dari hasil penyetaraan yang menggunakan metode RR dan RS. Dalam hal ini, metode SL dan HA satu level. Hasil penyetaraan yang menggunakan metode RR lebih baik dari hasil penyetaraan yang menggunakan metode RS. Terdapat 36 kombinasi peringkat level dari 72 kondisi yang ditinjau dengan kualifikasi sesuai dengan urutan nilai rata-rata RMSD kemampuan yang bersangkutan.

Daftar Pustaka

Arif Pratisto. (2004). *Cara mudah mengatasi masalah statistic dan rancangan percobaan dengan SPSS 12*. Jakarta: PT Elex Media Komputindo.

Bastari. (1998). *Comparison of IRT models that handle dichotomous and polytomous response data simultaneously*. Makalah, tidak diterbitkan, University of Massachusetts, Amherst.

Depdiknas. (2007). *Peraturan Menteri Pendidikan Nasional Republik Indonesia no. 20 Tahun 2007, tentang StandarPenilaian Pendidikan.*

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

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

Han, K.T., & Hambleton, R.K. (2007). *User's manual for wingen: Windows software that generates IRT model parameters and item responses*. Center for Educational Assessment Research Report. University of Massachusetts. Diambil pada tanggal 14 April 2007 dari http:// www.umass.edu/remp/software/wingen/

Hieronymus, A. N., Lindquistet, E. F., Hoover, H. D., et al. (1980). *Iowa test of basic for levels 7 & 8*. Iowa: The Riverside Publishing Company.

Jahja Umar. (1995). Berbagai permasalahan penggunaan bentuk soal uraian dan pilihan ganda dalam ujian. *Buletin Pengujian dan Penilaian, 6-10.*

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 transformation under unidimensional item response theory models*. Version 1.0 Iowa Testing Program. The University of Iowa. On line 8 Agustus 2006. http://www.uiowa.edu/Casma.

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

Mohandas, R. (2004). *Test equating*. Diambil pada tanggal 08 Januari 2007, dari http://www.Info worldbank. org/handout-equating.

Montgomery, D.C. (1984). *Design and analysis of experiments (2nd.ed.)*. New York: John Wiley & Sons.

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

Nana Suarna. (2005). *Pedoman panduan praktikum microsoft office EXCEL 2003*. Bandung: CV YRAMA WIDYA.

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

Sykes, R. C., & Yen, W. M. (2000). The scaling of mixed-item format test with the one-parameter and two-parameter partial credit model. *Journal of Educational Measurement, 37, 221-224.*

Biodata Penulis

Drs. Kartono, M.Si; Tempat dan Tanggal Lahir: Purwodadi Grobogan, 22 Februari 1956; Lulus S3 Program Studi Penelitian dan Evaluasi Pendidikan pada Tahun 2008. Pekerjaan: Dosen Jurusan Matematika FMIPA Universitas Negeri Semarang; Karya Ilmiah yang Relevan: 1) Kestabilan estimasi parameter butir tes dengan berbagai model IRT (2004) makalah tidak diterbitkan; 2) Analisis butir tes berbentuk uraian di bawah model GRM dan GPCM makalah disampaikan dalam seminar nasional pada tanggal 9 April 2005 di UPS Tegal; 3) Penskoran tes berbasis model respon butir, makalah disampaikan dalam seminar nasional pada tanggal 10 Desember 2005 di FMIPA UNNES; 4) Penyetaraan tes berbentuk pilihan ganda, makalah disampaikan dalam seminar nasional pada tanggal 27 Juli 2006 di Jurusan Matematika FMIPA UNNES; 5) Penyetaraan tes berbentuk uraian, makalah disampaikan dalam seminar nasional pada tanggal 24 November 2006 di FMIPA UNY.