Jurnal Evaluasi Pendidikan, Vol. 2, No. 1, Maret 2011, 101-114

METHOD OF SCORE EQUALITY AND SAMPLE SIZE

Tri Rijanto
Jurusan Teknik Elektro Fakultas Teknik, Universitas Negeri Surabaya (UNESA)
Jl. Ketintang, Surabaya
hari_tri2001@yahoo.com

Abstract

This study is aimed to obtain information of different score variance result of equating linear method and equipercentile method for sample size 200, 400, and 800 in Ujian Akhir Sekolah Berstandar Nasional (UASBN). The research is important in considering the test device of UASBN shaped packages of different tests. Scores obtained from different packages can not be directly inferred the existence of differences in ability between them, because the difficulty level of the package used influencing these differrences. To overcome the differences are doing through equating. The method used is an experiment of two variables, equating method and the number of respondents. The experiments are not conducted during the learning process, but conducted after the score and the pattern of the answers obtained through UASBN. The population examinee UASBN SD/MI 2008/2009 for IPA subject matter at East Jakarta. Sampling uses random replacement technique. The hypothesis is tested using similarity variance. The results with α = 0,05 shows: (1) score variance equipercentile method (σ^2_{ekp200}) is not different to score variance linear method (σ^2_{lin200}) for the sample size 200, (2) score variance equipercentile method (σ^2_{ekp400}) is not different to score variance equating linear method (σ^2_{lin400}) for the sample size 400, and (3) score variance equipercentile method (σ^2_{ekp800}) is not different to score variance equating linear method (σ^2_{lin800}) for the sample size 800.

***Keywords**: score variance, equating, equipercentile method, linear method*

METODE PENYETARAAN SKOR DAN UKURAN SAMPEL

Tri Rijanto
Jurusan Teknik Elektro Fakultas Teknik, Universitas Negeri Surabaya (UNESA)
Jl. Ketintang, Surabaya
hari_tri2001@yahoo.com

Abstrak

Penelitian ini bertujuan memperoleh informasi perbedaan variansi skor hasil penyetaraan metode linear dan metode ekipersentil untuk ukuran sampel 200, 400, dan 800 pada Ujian Akhir Sekolah Berstandar Nasional (UASBN). Penelitian ini penting dilakukan mengingat perangkat tes UASBN berbentuk paket tes yang berbeda. Skor yang diperoleh dari paket berbeda tidak dapat langsung disimpulkan adanya perbedaan kemampuan antarpeserta tes, karena tingkat kesukaran paket yang digunakan mempengaruhi perbedaan tersebut. Untuk menanggulangi perbedaan dilakukan melalui penyetaraan (*equating*). Metode yang digunakan adalah eksperimen dua variabel, yaitu metode penyetaraan dan ukuran sampel. Eksperimen tidak dilakukan selama proses pembelajaran berlangsung, tetapi dilakukan setelah skor dan pola jawaban peserta diperoleh melalui UASBN. Populasi penelitian peserta UASBN SD/MI tahun pelajaran 2008/2009 mata pelajaran IPA di Jakarta Timur. Penarikan sampel menggunakan teknik penarikan sampel acak dengan pengembalian. Hipotesis diuji menggunakan teknik analisis statistik uji kesamaan variansi. Hasil penelitian dengan $\alpha = 0{,}05$ menunjukkan: (1) variansi skor penyetaraan metode ekipersentil (σ^2_{ekp200}) tidak berbeda dengan variansi skor penyetaraan metode linear (σ^2_{lin200}) untuk ukuran sampel 200, (2) variansi skor penyetaraan metode ekipersentil (σ^2_{ekp400}) tidak berbeda dengan variansi skor penyetaraan metode linear (σ^2_{lin400}) untuk ukuran sampel 400, dan (3) variansi skor penyetaraan metode ekipersentil (σ^2_{ekp800}) tidak berbeda dengan variansi skor penyetaraan metode linear (σ^2_{lin800}) untuk ukuran sampel 800.

Kata kunci: variansi skor, *equating*, metode ekipersentil, metode linear

PENDAHULUAN

Ujian Nasional untuk Sekolah Dasar/Madrasah Ibtidaiyah/Sekolah Dasar Luar Biasa (SD/MI/SDLB) dilakukan pertama kali pada tahun 2008. Ujian tersebut bernama Ujian Akhir Sekolah Berstandar Nasional (UASBN) sesuai dengan Peraturan Menteri Pendidikan Nasional RI Nomor 39 tahun 2007. UASBN bertujuan untuk menilai pencapaian kompetensi lulusan secara nasional pada mata pelajaran bahasa indonesia, matematika, dan Ilmu Pengetahuan Alam (IPA) serta mendorong tercapainya target wajib belajar pendidikan dasar yang bermutu (Depdiknas, 2007: 6). UASBN SD/MI/SDLB dilaksanakan secara terintegrasi dengan ujian sekolah/ madrasah, artinya setiap paket soal UASBN terdiri dari 25% soal yang ditetapkan oleh Badan Standar Nasional Pendidikan (BSNP) dan berlaku secara nasional, serta 75% soal yang ditetapkan oleh penyelenggara UASBN tingkat provinsi berdasarkan spesifikasi yang ditetapkan

oleh BSNP. Dengan kata lain dalam paket soal terdiri dari 25% butir *anchor items* dan butir soal sisanya dibuat oleh masing-masing provinsi.

Instrumen penilaian yang digunakan oleh pemerintah dalam bentuk Ujian Nasional menurut Peraturan Menteri Pendidikan Nasional RI Nomor 20 tahun 2007 tentang Standar Penilaian Pendidikan memenuhi persyaratan substansi, konstruksi, bahasa, dan memiliki bukti validitas empirik serta menghasilkan skor yang dapat diperbandingkan antarsekolah, antardaerah, dan antartahun (Depdiknas, 2007: 9). Keterbandingan skor antarsekolah, kabupaten/kota, provinsi, dan antartahun dapat diperoleh jika semua peserta tes mengerjakan soal-soal (paket tes) yang sama. Perbedaan skor antarmereka menunjukkan perbedaan tingkat kemampuannya. Dalam praktiknya, pengadministrasian soal-soal sama antartahun, merugikan peserta tes yang mengerjakan pada tahun-tahun pertama dan menguntungkan mereka yang ikut tes pada tahun-tahun terakhir. Juga, pengadministrasian soal-soal yang sama di setiap sekolah sangat berisiko terhadap kebocoran.

Sebagai jalan keluar agar keadilan dan kerahasiaan soal-soal ujian terjaga mengharuskan pengadministrasian paket-paket tes berbeda antartahun, daerah, dan tempat tes. Tetapi masalah lain muncul, dengan mengadministrasian paket-paket yang berbeda, perbedaan skor antarpeserta tes tidak dapat langsung disimpulkan adanya perbedaan kemampuan antarmereka, karena tingkat kesukaran paket yang digunakan mempengaruhi perbedaan tersebut. Untuk menanggulangi ketidakadilan tersebut dilakukan penyamaan atau penyetaraan matriks skor *(equating)*. Penyetaraan matriks skor merupakan cara untuk memperoleh suatu konversi nilai dari skor mentah suatu paket, ke skor mentah paket yang lain. Jadi, melalui penyetaraan matriks skor dimungkinkan siswa menjawab benar 32 soal di Paket A, misalnya, mendapat nilai sama dengan siswa menjawab benar 30 soal di Paket B, karena Paket B lebih sukar dua soal dari Paket A. Akibat lebih serius dengan tidak adanya penyetaraan *(equating)* adalah ketidakadilan pada kelulusan. Batas lulus 5,25 misalnya, pada paket-paket sukar akan merugikan siswa-siswa yang mengerjakannya dan akan menguntungkan mereka yang mendapatkan paket-paket mudah. Oleh karena itu kemungkinan besar ada siswa yang seharusnya lulus tetapi karena mendapatkan paket sukar menjadi tidak lulus. Dengan demikian masalah penyetaraan menjadi penting untuk dikaji melalui penelitian yang komprehensif. Untuk dapat membandingkan atau menyetarakan skor Ujian Nasional diperlukan desain penyetaraan yang tepat. Menurut Hambleton dan Swaminathan (1985: 198) ada tiga desain dasar penyetaraan, yaitu metode kelompok tunggal *(single group method)*, metode kelompok ekuivalen *(equivalent group method)*, dan metode tes jangkar *(anchor test design).*

Terdapat dua cara penyamaan skor pada teori klasik yaitu penyamaan metode linear dan metode ekipersentil. Menurut Hambleton, Swaminathan, dan

Rogers (1991: 124) asumsi penyetaraan metode linear adalah kedua skor tes distribusinya berbeda, distribusi tersebut terkait dengan rerata dan simpangan bakunya. Dengan demikian pada metode linear membutuhkan asumsi rerata, dan simpangan baku, atau variansinya berbeda. Pada teori klasik, terdapat jumlah butir tertentu serta jumlah peserta tertentu. Dalam waktu yang bersamaan, semua peserta menjawab semua butir ujian. Jawaban setiap peserta terhadap setiap butir ujian dapat berupa jawaban benar atau berupa jawaban salah.

Pengukuran dalam pendidikan mengenal dua macam kekeliruan, yaitu kekeliruan acak atau kekeliruan sampel dan kekeliruan sistematik (Naga, 1992: 116). Kekeliruan sampel adalah perbedaan antara keadaan sebenarnya yang ada pada populasi. Hal ini disebabkan oleh hasil ukuran pada sampel tersebut hanya salah satu dari sekian banyak kemungkinan hasil pengukuran yang dapat dicuplik secara berulang-ulang dari suatu populasi. Kekeliruan sampel tetap saja muncul meskipun alat ukur yang dipakai, situasi dan kondisi pengukuran, maupun jenis kemampuan yang diukur tetap sama. Dengan demikian, ukuran sampel berpengaruh terhadap hasil pengukuran.

Resenfeld, dan Zirkel (1993: 156) menyatakan bahwa untuk sampel besar dengan ukuran sampel lebih dari 30 secara statistik akan menghasilkan suatu distribusi rerata cukup dekat pada distribusi normal, sehingga kalkulasi berdasarkan kurva normal masuk akal. Eid meneliti tentang pengaruh ukuran sampel pada penyetaraan butir tess mengusulkan untuk menggunakan ukuran sampel 200, 400, dan 800. Setiadi (1997: 7) dalam penelitiannya terhadap estimasi parameter butir menyatakan bahwa sampel yang relatif kecil berukuran 100 atau 200, sedangkan Livingston dan Feryok (1987: 9-10) melakukan penelitian pada penyetaraan ekipersentil estimasi frekuensi dengan penghalusan pada sampel berukuran 100 sampai dengan 3000 dan akurasi penyetaraan terjadi pada sampel berukuran 300. Penelitian ini menggunakan ukuran sampel 200, 400, dan 800.

Menurut Hambleton dan Swaminathan (1990: 199), penyetaraan skor bertujuan untuk mambandingkan skor yang diperoleh dari perangkat tes yang satu (X) dan perangkat tes lainnya (Y) yang dilakukan melalui proses penyetaraan skor pada kedua perangkat tersebut. Aiken (1997: 82) juga mengemukakan bahwa skor hasil pengukuran dua perangkat tes yang paralel dapat disetarakan untuk mengkonversi skor dari perangkat tes satu terhadap skor dari perangkat tes lainnya. Penyetaraan skor menurut Wiersma dan Jurs (1990: 148) adalah suatu prosedur empiris yang diperlukan untuk mentransformasikan skor suatu tes ke skor tes yang lain. Karena merupakan prosedur empiris maka penyetaraan skor didasarkan pada data skor tes. Demikian pula pendapat Kolen bahwa penyetaraan skor dapat dilakukan jika kelompok peserta setara, karena ketidaksetaraan yang ekstrim akan berpengaruh dalam perhitungan (Keeves, 1997: 733).

Penyetaraan ekipersentil menurut Braun & Holland dan Lord didasarkan pada definisi bahwa skala skor untuk kedua tes sebanding dan distribusi skor kedua tes identik (Linn, 1987: 247). Sementara itu menurut Croker dan Algina (1986: 346), secara umum diterima bahwa skor pada dua tes dianggap ekuivalen jika skor-skor tersebut memiliki tara persentil *(percentile rank)* yang sepadan. Menurut Braun dan Holland dan Lord penyetaraan ekipersentil didasarkan pada definisi bahwa skala skor untuk kedua tes sebanding dan distribusi skor ke dua tes identik (Linn, 1987: 247). Cook dan Petersen menyatakan penyetaraan ekipersentil tidak seperti metode penyetaraan lainnya, tidak membutuhkan asumsi-asumsi terhadap tes-tes yang akan disetarakan (Naga, 1992: 226). Dengan demikian metode penyetaraan ekipersentil mengasumsikan bahwa skor pada tes X dan Y adalah identik atau ekivalen, bertujuan agar distribusi dari skor tes X yang diubah sama dengan distribusi skor tes Y, dan tidak membutuhkan asumsi-asumsi seperti pada metode linear. Menurut Naga (1992: 364), penyamaan ekipersentil semata-mata hanya melihat ke persentil skor yang disetarakan, tanpa memperhatikan bentuk distribusi probabilitas dari kedua skor yang akan disetarakan matriknya. Jika bentuk distribusi probabilitas kedua skor tersebut sama, maka cara penyamaan ini cukup baik. Akan tetapi jika bentuk distribusi probabilitas kedua skor tidak sama, maka metode ekipersentil akan menyamakan persentil pada bentuk distribusi probabilitas yang berbeda. Dengan kata lain penyetaraan ekipersentil tidak perlu memperhatikan distribusi probabilitas kedua skor yang akan disetarakan matriknya.

Oleh karena itu, penelitian ini bertujuan untuk mengetahui perbedaan variansi skor hasil penyetaraan antara metode menyetaraan linear dan metode penyetaraan ekipersentil untuk ukuran sampel 200, 400 dan 800.

METODE PENELITIAN

Metode penelitian ini adalah metode eksperimen yang terdiri dari dua variabel. Eksperimen tidak dilakukan selama proses pembelajaran berlangsung, akan tetapi dilakukan setelah skor dan pola jawaban peserta tes diperoleh melalui pelaksanaan UASBN SD/MI negeri dan swasta tahun pelajaran 2008/2009 mata pelajaran IPA. Variabel bebas dalam penelitian ini adalah metode penyetaraan dan ukuran sampel. Metode penyetaraan pada penelitian ini adalah metode penyetaraan linear dan metode penyetaraan ekipersentil, sedangkan variabel ukuran sampel berturut-turut adalah 200, 400, dan 800. Variabel terikatnya adalah variansi skor hasil penyetaraan.

Populasi penelitian dibagi menjadi dua jenis yaitu populasi peserta tes dan populasi skor responden. Populasi peserta tes adalah peserta tes UASBN tahun pelajaran 2008/2009 mata pelajaran IPA se-Jakarta Timur. Populasi tersebut sebanyak 44.401 siswa yang terbagi menjadi dua bagian yaitu sebanyak 22.201 siswa yang mengerjakan paket 01 dan sebanyak 22.200 siswa yang mengerjakan paket 02. Pengambilan sampel dengan pengembalian metode acak berulang dengan pengembalian.

Data diperoleh melalui Pusat Penilaian Pendidikan Badan Penelitian dan Pengembangan Kementerian Pendidikan Nasional. Data tersebut berupa matriks jawaban siswa dan kunci jawaban dalam format *notepad.* Oleh karena itu agar dapat dianalisis menggunakan MINITAB, *ITEMAN* dan Excel, data tersebut harus diverifikasi terlebih dahulu. Teknik analisis data menggunakan uji kesamaan variansi menggunakan analisis Uji F. Desain penelitian ini dapat dilihat pada tabel 1.

Tabel 1. Desain Penelitian

Jumlah Sampel	Metode Penyetaraan	
	Metode Linear	Metode Ekipersentil
n = 200	S^2*lin200*	S^2*ekp200*
n = 400	S^2*lin400*	S^2*ekp 400*
n = 800	S^2*lin800*	S^2*ekp 800*

HASIL PENELITIAN

Perangkat IPA Paket 01

Parameter pertama yaitu indeks kesukaran butir, baik *anchor items* maupun butir daerah. Deskripsi indeks kesukaran butir dapat dirangkum pada tabel 2. Tabel tersebut menunjukkan bahwa kategori butir pada ukuran sampel 200, 400, dan 800 tidak berbeda. Butir-butir tersebut mengelompok pada kategori butir yang sama, baik untuk butir pusat maupun butir daerah. Hal ini menunjukkan bahwa ukuran sampel tidak mempengaruhi indeks kesukaran butir.

Tabel 2. Rangkuman Indeks Kesukaran Butir

Ukuran sampel	**Kategori Butir**	**Asal dan Nomor Butir**		**Jumlah (%)**	**Total**
		Pusat	**Daerah**		
200	Mudah	1, 2, 30, 40	3, 5, 6, 7, 10, 12,18, 21, 25, 26, 27,29, 31, 34, 38	19 (47,5%)	40 (100%)
	Sedang	9, 20, 22, 24	4, 8, 11, 13, 15, 17, 19, 28, 32, 35, 36, 37, 39	17 (42,5%)	
	Sukar	16, 23	14, 33	4 (10%)	
400	Mudah	1, 2, 30, 40	3, 5, 6, 7, 10, 12, 18, 21, 25, 26, 27, 29, 31, 34, 38	19 (47,5%)	40 (100%)
	Sedang	9, 20, 22, 24	4, 8, 11, 13, 15, 17, 19, 28, 32, 35, 36, 37, 39	17 (42,5%)	
	Sukar	16, 23	14, 33	4 (10%)	
800	Mudah	1, 2, 30, 40	3, 5, 6, 7, 10, 12, 18, 21, 25, 26, 27, 29, 31, 34, 38	19 (47,5%)	40 (100%)

Ukuran sampel	Kategori Butir	Asal dan Nomor Butir		Jumlah (%)	Total
		Pusat	Daerah		
	Sedang	9, 20, 22, 24	4, 8, 11, 13, 15, 17, 19, 28, 32, 35, 36, 37, 39	17 (42,5%)	
	Sukar	16, 23	14, 33	4 (10%)	

Parameter berikutnya adalah daya beda butir. Dilihat dari kategori daya beda butir sebanyak 30 butir termasuk memadai dan sebanyak 10 butir tidak memadai. Dilihat dari ukuran sampel butir-butir tersebut mengelompok pada kategori yang sama. Hanya satu butir *anchor items* yang tidak memadai pada analisis ukuran sampel 800. Dengan demikian dapat dikatakan bahwa ukuran sampel tidak banyak mempengaruhi terhadap daya beda butir. Rangkuman selengkapnya daya beda butir dapat dilihat pada tabel 2.

Parameter selanjutnya adalah reliabilitas tes. Dengan sepuluh kali replikasi diperoleh rerata reliabilitas tes yang memadai. Dilihat dari jumlah pengambilan sampel rerata reliabilitas tes relatif sama yaitu termasuk reliabilitas tes yang memadai. Jadi, penarikan ukuran sampel 200, 400, dan 800 tidak mempengaruhi terhadap besarnya rerata reliabilitas tes.

Perangkat IPA Paket 02

Dilihat dari parameter indeks kesukaran butir, menunjukkan bahwa kategori butir pada analisis ukuran sampel 200 dan 400 tidak berbeda. Pada analisis ukuran sampel 800, butir nomor 40 (butir pusat) dan 18 (butir daerah) pada analisis dengan ukuran sampel 200, dan 400 dalam kategori mudah, sedangkan pada analisis ukuran sampel 800 dalam kategori sedang. Dengan demikian secara keseluruhan butir-butir tersebut mengelompok pada kategori butir yang sama, baik untuk butir pusat maupun butir daerah. Hal ini menunjukkan ukuran sampel 200 dan 400 tidak mempengaruhi kategori indeks kesukaran butir, sedangkan ukuran sampel 800 mempengaruhi dua butir berubah dari butir kategori mudah menjadi kategori sedang. Rangkuman indeks kesukaran butir dapat dilihat pada tabel 3.

Tabel 3. Rangkuman Indeks Kesukaran Butir

Ukuran sampel	Kategori Butir	Asal dan Nomor Butir		Jumlah (%)	Total (%)
		Pusat	Daerah		
200	Mudah	1, 2, 30, 40	3, 4, 5, 6, 7, 8, 12, 18, 21, 25, 26, 27, 28, 33, 34, 35, 39	21 (52,5%)	40 (100%)
	Sedang	9, 20, 22, 24	10, 11, 13, 14, 17, 19, 29, 31, 32, 36	14 (35%)	
	Sukar	16, 23	15, 37, 38	5 (12,5%)	

Ukuran sampel	Kategori Butir	Asal dan Nomor Butir		Jumlah (%)	Total (%)
		Pusat	Daerah		
400	Mudah	1, 2, 30, 40	3, 4, 5, 6, 7, 8, 12, 18, 21, 25, 26, 27, 28, 33, 34, 35, 39	21 (52,5%)	40 (100%)
	Sedang		10, 11, 13, 14, 17, 19, 29, 31, 32, 36	14 (35%)	
	Sukar	16, 23	15, 37, 38	5 (12,5%)	
800	Mudah	1, 2, 30	3, 4, 5, 6, 7, 8, 12, 21, 25, 26, 27, 28, 33, 34, 35, 39	19 (47,5%)	40 (100%)
	Sedang	9, 20, 22, 24, 40	10, 11, 13, 14, 17, 18, 19, 29, 31, 32, 36	16 (40%)	
	Sukar	16, 23	15, 37, 38	5 (10,5%)	

Dilihat dari persentase butir mudah, sedang dan sukar, pada analisis ukuran sampel 200 dan 400, sebesar 52,5% butir tergolong mudah, dan pada ukuran sampel 800 sebesar 47,5% butir tergolong mudah. Dengan kata lain perangkat tes IPA Paket 02 separuh butir tergolong mudah, sehingga memberi peluang kepada peserta tes untuk lulus dengan standar minimal.

Tabel 4. Rangkuman Daya Beda Butir

Ukuran sampel	Kategori Daya Beda	Nomor Butir		Jumlah (%)	Total (%)
		Pusat	Daerah		
200	Memadai	1, 2, 9, 16, 20,22,24,40	3, 7, 8, 10, 13, 15, 18, 19, 21, 25, 26, 27, 28, 29, 31, 32, 34, 36, 39	27 (67,5%)	40 (100%)
	Tidak Memadai	23,30	4, 5, 6, 11, 12, 14, 17, 33, 35, 37, 38	13 (32,5%)	
400	Memadai	1, 2, 9, 16, 20,22,24,40	3, 7, 8, 10, 13, 15, 18, 19, 21, 25, 26, 27, 28, 29, 31, 32, 34, 36, 39	27 (67,5%)	40 (100%)
	Tidak Memadai	23,30	4, 5, 6, 11, 12, 14, 17, 33, 35, 37, 38	13 (32,5%)	
800	Memadai	1, 2, 9, 16, 20, 22, 23,	3, 7, 8, 10, 13, 15, 18, 19, 21, 25, 26, 27, 28,	28 (70%)	40 (100%)

Ukuran sampel	Kategori Daya Beda	Nomor Butir		Jumlah (%)	Total (%)
		Pusat	Daerah		
		24,40	29, 31, 32, 34, 36, 39		
	Tidak Memadai	30	4, 5, 6, 11, 12, 14, 17, 33, 35, 37, 38	12(30%)	

Dilihat dari parameter daya beda butir sebanyak 27 butir termasuk memadai dan sebanyak 13 butir tidak memadai. Ini terjadi pada analisis ukuran sampel 200 dan 400 yang mempunyai daya beda, kategori butir, dan mengelompok pada kategori yang sama. Pada analisis ukuran sampel 800, satu butir yang tadinya masuk dalam kategori tidak memadai menjadi memadai, yaitu butir nomor 23 yang berasal dari butir *anchor items*. Hanya satu butir *anchor items* yang tidak memadai pada analisis ukuran sampel 800. Dengan demikian dapat dikatakan bahwa ukuran sampel tidak banyak mempengaruhi terhadap daya beda butir. Rangkuman selengkapnya daya beda butir dapat dilihat pada tabel 3.

Dilihat dari parameter reliabilitas tes dengan sepuluh kali replikasi diperoleh rerata reliabilitas tes yang memadai. Dilihat dari jumlah pengambilan sampel rerata reliabilitas tes relatif sama yaitu termasuk reliabilitas tes yang memadai. Penarikan ukuran sampel 200, 400, dan 800 tidak mempengaruhi terhadap besarnya rerata reliabilitas tes.

Variansi Skor Hasil Penyetaraan

Analisis variansi skor hasil penyetaraan dilakukan terhadap masing-masing cuplikan dengan ukuran sampel 200, 400, dan 800 baik untuk paket 01 maupun paket 02, masing-masing sebanyak 41 kali cuplikan. Jadi, seluruhnya terdapat 246 analisis dan menghasilkan variansi skor hasil penyetaraan dengan metode linear dan metode ekipersentil. Rerata variansi skor untuk s^2_{lin200} sebesar 25,119, s^2_{lin400} sebesar 24,710, dan s^2_{lin800} sebesar 24,885. Simpangan bakunya berturut-turut adalah untuk s^2_{lin200} sebesar 2,241, s^2_{lin400} sebesar 2,591, dan s^2_{lin800} sebesar 1,971. Dengan demikian dapat dikatakan rerata variansi untuk sampel 200, 400, dan 800 tidak jauh berbeda.

Selanjutnya hasil analisis menggunakan metode ekipersentil rerata variansi skor untuk sampel 200 sebesar 29,506, sampel 400 sebesar 29,572, dan sampel 800 sebesar 29,132. Simpangan bakunya berturut-turut adalah sampel 200 sebesar 2,124, sampel 400 sebesar 1,831, dan sampel 800 sebesar 1,599. Dengan demikian dapat dikatakan bahwa rerata variansi skor ketiga model pensampelan tidak jauh berbeda.

PENGUJIAN HIPOTESIS

Pengujian Hipotesis Pertama

Hipotesis pertama dinyatakan bahwa variansi skor hasil penyetaraan menggunakan metode ekipersentil *(s^2_{ekp200})* lebih besar daripada metode penyetaraan linear *(s^2_{lin200})* untuk ukuran sampel 200. Rerata variansi skor hasil penyetaraan yang dihasilkan dari penyetaraan menggunakan metode ekipersentil (s^2_{ekp200}) adalah 4,468. Kemudian besarnya rerata variansi skor hasil penyetaraan yang dihasilkan menggunakan metode linear (s^2_{lin200}) sebesar 5,218. Untuk memastikan apakah kedua variansi skor ini berbeda secara signifikan atau tidak, maka dilakukan analisis uji-F sebagai berikut: $F_{h1} = s^2_{ekp200} / s^2_{lin200} = 0{,}856$.

Harga F_{tabel} untuk taraf signifikansi (α) = 0,05 dengan derajat pembilang = 40 dan penyebut = 40 sebesar 1,684. Nilai F_{h1} yang diperoleh lebih kecil dibandingkan dengan F_{tabel}, dengan demikian hipotesis nol diterima. Artinya, variansi skor hasil penyetaraan menggunakan metode ekipersentil (s^2_{ekp200}) tidak berbeda dengan variansi skor hasil penyetaraan menggunakan metode linear (s^2_{lin200}) untuk ukuran sampel 200.

Pengujian Hipotesis Kedua

Hipotesis kedua dinyatakan bahwa variansi skor hasil penyetaraan menggunakan metode ekipersentil *(s^2_{ekp400})* lebih besar daripada metode penyetaraan linear *(s^2_{lin400})* untuk ukuran sampel 400. Rerata variansi skor hasil penyetaraan yang dihasilkan dari penyetaraan menggunakan metode ekipersentil (s^2_{ekp400}) adalah 3,275. Kemudian besarnya rerata variansi skor hasil penyetaraan yang dihasilkan menggunakan metode linear (s^2_{lin400}) sebesar 6,684. Untuk memastikan apakah kedua variansi skor ini berbeda secara signifikan atau tidak, maka dilakukan analisis uji-F sebagai berikut: $F_{h2} = s^2_{ekp400} / s^2_{lin400} = 0{,}490$.

Harga F_{tabel} untuk taraf signifikansi (α) = 0,05 dengan pembilang = 40 dan penyebut = 40 sebesar 1,684. Nilai F_{h2} yang diperoleh lebih kecil dibandingkan dengan F_{tabel}, dengan demikian hipotesis nol diterima. Artinya, variansi skor hasil penyetaraan menggunakan metode ekipersentil (s^2_{ekp400}) tidak berbeda dengan variansi skor hasil penyetaraan menggunakan metode linear (s^2_{lin400}) untuk ukuran sampel 400.

Pengujian Hipotesis Ketiga

Hipotesis ketiga dinyatakan bahwa variansi skor hasil penyetaraan menggunakan metode ekipersentil *(s^2_{ekp800})* lebih besar daripada metode penyetaraan linear *(s^2_{lin800})* untuk ukuran sampel 800. Rerata variansi skor hasil penyetaraan yang dihasilkan dari penyetaraan menggunakan metode ekipersentil (s^2_{ekp800}) adalah 2,575. Kemudian besarnya rerata variansi skor hasil penyetaraan yang dihasilkan menggunakan metode linear (s^2_{lin800}) sebesar 3,878. Untuk memastikan apakah kedua variansi skor ini berbeda secara signifikan atau tidak, maka dilakukan analisis uji-F sebagai berikut: $F_{h3} = s^2_{ekp800} / s^2_{lin800} = 0{,}664$.

Harga F_{tabel} untuk taraf signifikansi (α) = 0,05 dengan pembilang = 40 dan penyebut = 40 sebesar 1,684. Angka F_{h3} yang diperoleh lebih kecil dibandingkan dengan F_{tabel} jadi, hipotesis nol diterima. Artinya, variansi skor hasil penyetaraan menggunakan metode ekipersentil (s^2_{ekp800}) tidak berbeda dengan variansi skor hasil penyetaraan menggunakan metode linear (s^2_{lin800}) untuk ukuran sampel 800.

Hasil pengujian hipotesis pertama di atas mengungkapkan bahwa variansi skor penyetaraan (s^2_{lin200}) menggunakan metode linear mempunyai variansi yang sama dengan variansi skor penyetaraan (s^2_{ekp200}) menggunakan metode ekipersentil dengan ukuran sampel 200. Kemudian hasil pengujian hipotesis kedua diperoleh kenyataan bahwa variansi skor penyetaraan (s^2_{lin400}) menggunakan metode linear juga mempunyai variansi yang sama dengan variansi skor penyetaraan (s^2_{ekp400}) menggunakan metode ekipersentil dengan ukuran sampel 400. Demikian pula untuk pengujian hipotesis ketiga mengungkapkan bahwa variansi skor penyetaraan (s^2_{lin800}) menggunakan metode linear mempunyai variansi yang sama dengan variansi skor penyetaraan (s^2_{ekp800}) menggunakan metode ekipersentil dengan ukuran sampel 800. Kenyataan ini menunjukkan bahwa variansi skor penyetaraan kedua metode tersebut mempunyai variansi skor yang tidak berbeda untuk ukuran sampel 200, 400, dan 800 dengan replikasi 41 kali.

PEMBAHASAN

Berdasarkan hasil pengujian hipotesis pertama, kedua, dan ketiga diperoleh tidak ada perbedaan variansi antara metode linear dan ekipersentil dengan ukuran sampel 200, 400, dan 800. Hal ini disebabkan oleh berbagai kemungkinan, antara lain adalah asumsi penggunaan penyetaraan metode linear tidak terpenuhi. Hal ini sesuai dengan pendapat Hambleton dan Rogers bahwa asumsi penyetaraan dengan cara linear adalah kedua skor tes distribusinya berbeda, distribusi tersebut terkait dengan rerata dan simpangan bakunya. Jika terdapat kasus yang demikian skor kedua kelompok akan sama. Sebaliknya bila asumsi tersebut benar, maka penyetaraan linear menjadi penyetaraan ekipersentil (Hambleton, Swaminathan, dan Rogers, 1991: 124-125). Dengan demikian distribusi skor kedua kelompok pada penelitian mempunyai distribusi yang tidak berbeda, sehingga penyetaraan linear sama dengan penyetaraan ekipersentil. Dalam penelitian ini tidak dilakukan pemeriksaan atau pengujian terhadap distribusi kedua skor tes terkait dengan rerata dan simpangan bakunya. Oleh karena itu terdapat kemungkinan besar kedua distribusi kedua skor tidak berbeda sehingga penyetaraan linear menjadi penyetaraan ekipersentil. Dengan kata lain metode penyetaraan linear sama dengan metode penyetaraan ekipersentil jika kedua distribusi skor tidak ada berbeda.

Diterimanya hipotesis nol dalam penelitian ini selain tidak dilakukan pemeriksaan atau pengujian terhadap distribusi kedua skor tes terkait dengan

rerata dan simpangan baku, juga tidak dilakukan pemeriksaan atau pengujian terhadap reliabilitas kedua tes secara statistik. Hal ini sesuai dengan pendapat Lord bahwa skor mentah pada tes yang mempunyai reliabilitas tidak sama tidak dapat disetarakan (sebaliknya ketika skor dari tes yang tidak reliabel dapat disetarakan terhadap skor pada tes yang reliabel, akan meniadakan kebutuhan konstruksi tes yang reliabel) (Hambleton, Swaminathan, dan Rogers, 1991: 124-125). Dengan demikian reliabilitas perlu dilakukan pengujian secara statistik terhadap reliabilitas kedua tes untuk mendapatkan informasi kesamaan reliabilitasnya.

Di samping itu dalam penelitian ini juga tidak dilakukan pengujian secara statistik terhadap taraf kesukaran butir kedua tes. Menurut Lord skor mentah pada tes dengan taraf kesukaran yang bervariasi tidak dapat disetarakan karena perangkat tes tidak akan sama reliabilitasnya pada tingkatan kemampuan yang berbeda (Hambleton, Swaminathan, dan Rogers, 1991: 124-125). Jadi, perlu dilakukan pengujian secara statistik terhadap taraf kesukaran kedua tes untuk mengetahui kesamaan taraf kesukaran kedua tes terutama untuk butir gandeng *(anchor items).*

SIMPULAN

Metode penyetaraan linear dan ekipersentil dapat dimanfaatkan untuk membuat peta mutu pendidikan atau hasil belajar peserta didik dalam suatu wilayah tertentu. Hasil belajar yang diperoleh melalui UASBN dapat ditampilkan dalam skor penyetaraan dari kelompok (paket) ke kelompok (paket) lain yang setara. Penggunaan metode linear mengharuskan adanya uji distribusi skor terkait dengan rerata dan simpangan bakunya. Jika distribusi skor (rerata dan simpangan baku) sama, maka dapat menggunakan salah satu metode penyetaraan. Penggunaan kedua metode tersebut tidak dipengaruhi oleh ukuran sampel sesuai dengan hasil penelitian ini. Namun demikian sebelum menggunakan salah satu metode tersebut perlu dilakukan uji reliabilitas perangkat tes. Jika kedua perangkat mempunyai reliabilitas sama maka penyetaraan dapat dilanjutkan. Penggunaan salah satu metode penyetaraan mempertimbangkan taraf kesukaran tes. Taraf kesukaran yang bervariasi tidak dapat disetarakan karena perangkat tes tidak akan sama reliabilitasnya pada tingkatan kemampuan yang berbeda. Informasi ini penting agar dalam menggunaan atau memilih salah satu metode melihat variasi taraf kesukaran butir tes. Akan menimbulkan hasil yang tidak akurat jika asumsi tersebut tidak dipenuhi dalam menggunakan salah satu metode penyetaraan.

DAFTAR PUSTAKA

Aiken, Lewis R. (1997). *Psychological Testing and Assemesment.* Boston: Allyn & Bacon.

Brennan, R. L., dan Michael J. Kolen. (1991). "Some Practical Issues in Equating." *Applied Psychological Measurement*, Vol. 28 (3).

Brown, Frederick G. (1976). *Principles of Educational and Psychologycal Testing*. New York: Holt, Rinehart, and Winston.

Croker, Linda dan James Algina. (1986). *Introduction to Classical & Modern Test Theory.* New York: Rinehart and Winston Inc.

Eid, Ghada K. "*Effects of Sample Size in the Equating of Test Items*" http://findarticles.com/p/articles/mi_qa3673/is_200510/ai_n15641924 (diakses tanggal 12 Oktober 2008).

Hambleton, Ronald K., dan Hariharan Swaminathan. (1985). *Item Response Theory: Prin-ciples and Application.* Boston: Kluwer.

_____. (1990). *Item Respon Theory Principles and Applications.* Boston: Kluwer Nijhoff Publishing.

Hambleton, Ronald K., (1991). Hariharan Swaminathan, dan H. Jane Rogers. *Fundamentals of Item Respons Theory*. California: SAGE Publications, Inc.

Keeves, John (ed). 1997). *An International Hand Book of Measurement*. Oxford: Pergamon.

Kolen, Michael J., dan Robert L. Brennan. (1995). *Test Equating Methods and Practices.* New York: Springer.

Kolen, Michael J., dan Douglas R. Whitney. (1982). "Comparison of Four Prosedur for Equating the Test of General Educational Development." *Journal of Educa-tional Measurement*, Vol. 10 (4).

Linn, Robert L. (1987). *Educational Measurement*. New York: Macmillan Publishing Company.

Livingston, S. A., N. J. Doran, dan N. K. Wright. (1990). "What Combination of Sampling and Equating Methods Work Best?." *Applied Measurement in Education*, Vol. 3.

Lord, Frederick M. dan Melvin R. Novick. (1968). *Statistical Theories of Mental Test Scores.* Massachusetts: Addison-Wesley Publishing Company, Inc.

Marco, Gory L. *et al*. (1993). *A Test of The Adequacy of Curvilinear Score Equating Models New Horizon in Testing*. New York: Academic Press.

Montgomery, Dougals C. (2001). *Design and Analysis of Experiments.* New Jersey: John Willey & Sons, Inc.

Naga, Dali S. (1992). *Pengantar Teori Sekor.* Jakarta: Besbats.

Naiman, Arnold, Robert Rosenfeld, dan Gene Zirkel. (1993). *Understanding Statistic.* Singapore: McGraw-Hill.

Thorndike, Robert M. *et al.* (1991). *Measurement and Evaluation in Psychology and Education.* New York: Macmillan Publishing Company.

Wiersma, William dan Stephen G. Jurs. (1990). *Educational Measurement and Testing*. Boston: Allyn and Bacon.

Departemen Pendidikan Nasional. Peraturan Pemerintah Republik Indonesia Nomor 19 tahun 2005 tentang Standar Nasional Pendidikan. Jakarta: Badan Standar Nasional Pendidikan.

Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 39 tahun 2007 tentang Ujian Akhir sekolah Berstandar Nasional (UASBN) untuk Sekolah Dasar/Madrasah Ibtidaiyah/Sekolah Luar Biasa (SD/MI/SDLB) tahun Pelajaran 2007/2008. Jakarta: Badan Standar Nasional Pendidikan, 2007.