PENYETARAAN TES UAN : MENGAPA DAN BAGAIMANA?

Sukirno DS
FISE Universitas Negeri Yogyakarta

Abstract

Debates concerning what is right and what is wrong with national final examination (UAN) in Indonesia are not new. National Final Examination (UAN), a significant process using a test form to measure learning output, has been developed to provide useful information for the decision makers (parents, educators, policy-makers and the local community). Multiple forms of a certification exam are desirable for a variety of reasons. However, the problem of comparability among test scores using different test forms must be addressed in order to insure fairness and consistency in each testing situation. Test forms must be interchangeable across test administrations. Psychometric procedures known as equating methods can be utilized to produce comparable (equated) scores. Equating procedures consist of (1) a design for collecting test data for equating, (2) a clearly defined level of expected correspondence among test scores, and (3) specific statistical procedures that are used to estimate score correspondence. The process of equating is used to obtain comparable scores when more than one test forms are used in a test administration. In many situations in test administration, more than one form of the tests are used for security reasons. Beside, by test equating, a test form can be administered more flexible in the context of environment and time. There are several techniques and methodologies that can be used in equating test forms. Generally speaking, these techniques and methodologies can be divided into three major activities, namely determining test equating design, determining the test equating methods, and determining the way of test equating will be taken. The magnitude of standard error of estimate (SEE) is used to evaluate which is the most accurate method of test equating. The less the score of standard error of estimate the more accurate of the test equating method.

Keywords : penyetaraan tes, UAN

A. Pendahuluan

Perdebatan tentang UAN muncul tidak hanya karena kebijakan UAN yang digulirkan Departemen Pendidikan Nasional minim sosialisasi dan tertutup, tapi lebih pada hal yang bersifat fundamental secara yuridis dan pedagogis. Pada aspek pedagogis, dalam ilmu kependidikan, kemampuan peserta didik mencakup tiga aspek, yakni pengetahuan (kognitif), keterampilan (psikomotorik), dan sikap (afektif), tetapi yang dinilai dalam UAN hanya satu aspek kemampuan, yaitu kognitif, sedangkan kedua aspek lain tidak diujikan sebagai penentu kelulusan.

Sedangkandari aspek yuridis, beberapa pasal dalam UU Sistem Pendidikan Nasional Nomor 20 Tahun 2003 telah dilanggar, misalnya pasal 35 ayat 1 yang menyatakan bahwa standar

(jumlah soal yang dijawab dengan benar) siswa yang diperoleh dari paket tes yang berbeda tingkat kesukarannya telah dilakukan penyesuaian. Dengan begitu, tabel konversi untuk masing - masing paket tes adalah setara. Skala baku nasional memungkinkan dilakukannya analisis perbandingan kemampuan (mutu *outcome*) antara sekolah, antardaerah, antarwilayah di mana paket tes digunakan berbeda. Selain itu, akan memungkinkan pula dilakukannya pemantauan mutu pendidikan secara berkesinambungan dari tahun ke tahun.

Informasi hasil UAN antartahun, antarsekolah, antardaerah, dan antarwilayah dapat diperbandingkan (komparabel) sehingga dapat digunakan dalam rangka mengendalikan mutu pendidikan itu sendiri, sekaligus merumuskan kebijakan dalam rangka peningkatan mutu pendidikan secara nasional. Konversi juga dianggap memudahkan ketafsiran dari hasil UAN. Dengan penafsiran tersebut, nilai UAN dapat memberikan informasi tentang apa yang siswa kuasai dan apa yang tidak dikuasainya sesuai kompetensi dan tujuan pembelajaran setiap bidang sesuai kurikulum yang berlaku.

Ujian Akhir Nasional untuk pengendalian mutu tetap diperlukan dan merupakan kapasitas pemerintah (pusat) untuk mengukur, kompetensi siswa saat ini, seberapa jauh jaraknya terhadap standar nasional, dan membandingkan kompetensi itu antarsekolah, antardaerah, serta antarwaktu. UAN juga berguna untuk sistem penjaminan mutu, dengan menggunakan hasil UAN dapat dimonitor dan ditelaah kapasitas guru-guru, fasilitas pendidikan, serta proses pembelajaran agar dapat diketahui setiap saat apa yang harus dilakukan pemerintah atau sekolah agar mutu pendidikan secara konstan meningkat (Ace Suryadi, 2007).

Ujian Akhir Nasional (UAN) sebagai salah satu proses pengukuran hasil belajar tingkat nasional, memiliki tujuan dan kegunaan yang penting dalam bidang pendidikan. Hasil UAN akan digunakan sebagai dasar pengambilan berbagai keputusan strategis di bidang pendidikan. Keputusan strategis tersebut, antara lain: 1) untuk mengetahui sejauh mana tujuan kurikulum telah tercapai, 2) sebagai sarana dalam pemantauan dan penentuan standarisasi mutu pendidikan nasional, 3) sebagai bahan acuan dan pertimbangan dalam menentukan kelulusan, dan 4) sebagai perangkat seleksi bagi penerimaan siswa baru ke jenjang pendidikan berikutnya (Depdikbud, 1998). Ditinjau dari fungsinya, hasil UAN selain digunakan untuk memantau mutu pendidikan juga sebagai penentu kelulusan seseorang untuk seleksi masuk jenjang pendidikan berikutnya. Menurut Dananwijaya (2000), adanya beberapa tujuan yang ingin dicapai dari pelaksanaan UAN sering menimbulkan masalah.

Terlepas dari pro dan kontra tentang UAN, informasi yang diperoleh melalui UAN yang terdiri lebih dari satu paket, tes harus benar-benar mencerminkan kemampuan peserta UAN yang sebenarnya. Artinya perbedaan skor yang diperoleh seorang peserta lainnya adalah semata-mata karena perbedaan kemampuan di antara mereka, bukan disebabkan karena faktor lain, misalnya karena mengerjakan paket tes UAN sebagai alat ukur mempunyai kualitas yang baik dan digunakan sesuai dengan prosedur penilaian yang benar dan hati-hati, serta adanya bobot penyetaraan antar paket tes (Supriyoko, 2000).

B. Pembahasan

1. Penyetaraan Tes

Pusat Penelitian dan Pengembangan Sistem Pengujian (Puslitbang Sisjian) sudah mengembangkan sistem pengujian dengan pembentukan bank tes untuk SLTP dan SMU yang dikalibrasi. Proses kalibrasi dilakukan dengan pendekatan model Logistik satu parameter. Paket tes disusun dari butir tes yang ada di bank tes tersebut. Menurut Hayat (1995), penyetaraan antarpaket tes dilakukan secara otomatis dengan komputer Program *Bigstep* tanpa metode penyetaraan. Hal ini dilakukan karena cakupan UAN yang berskala nasional, waktu analisis yang relatif singkat, dan pertimbangan segi kepraktisan dari model logistik satu parameter. Penyetaraan paket tes dilakukan menyetarakan isi bukan item tes, sehingga bisa jadi item tes yang digunakan memiliki tingkat kesulitan berbeda tergantung kepada cara pengembang tes mengemas soal.

Meskipun paket tes UAN disusun dari bank tes yang sudah dikalibrasi, namun kesetaraan antarpaket tes masih harus diperhatikan. Hal ini dilakukan karena proses pembuatan bank tes berdasarkan hasil estimasi sehingga tidak terlepas dari kesalahan pengestimasian walaupun kecil. Kesalahan kecil ini jika tidak diperhatikan akan terakumulasi, akibat proses kalibrasi yang terus menerus dari waktu ke waktu (Wright dan Stone, 1979). Menurut Petersen, Kolen, dan Hoover (1989), hasil pengukuran menjadi kurang tepat, karena adanya kesalahan pengukuran. Oleh karena itu, kesalahan pengukuran ini mengakibatkan konstanta konversi antarpaket tes UAN yang diestimasi kurang tepat.

Hayat dan Pranata (1995) menyatakan bahwa tanpa prosedur penyetaraan dari paket tes atau perangkat tes yang berbeda akan tedapat beberapa kelemahan, antara lain: 1) nilai keterbandingan hasil tes bagi siswa atau sekolah yang mengambil perangkat tes yang berbeda akan beerkurang, 2) tidak adanya jaminan bahwa perangkat tes yang dikembangkan dengan kisi-kisi yang sama mempunyai tingkat kesulitan yang sama, dan 3) karena tidak adanya nilai keterbandingan hasil tes mengakibatkan informasi yang akurat tentang pencapaian mutu belajar siswa tidak mudah untuk diperoleh.

Hambleton dan Swaminathan (1985) dan Stroud dalam Holland & Rumib (1982) berpendapat bahwa dengan adanya perbedaan paket tes, ada kemungkinan perbedaan dalam hal: karakteristik, sifat dan kemampuan, dan tingkat kesulitan yang diukur. Memperhatikan pendapat-pendapat tersebut, agar tingkat kemampuan siswa dan kualitas pembelajaran dapat dibandingkan maka tes yang terdiri lebih dari satu paket tes perlu disetarakan.

Hembleton dan Swiminathan (1985) menegaskan bahwa, sekalipun perangkat tes disusun berdasarkan kisi-kisi yang sama, namun jarang sekali atau hampir tidak pernah ditemukan perangkat tes yang benar-benar setara dalam sebaran serta tingkat kesukaran. Pendapat lain dikemukakan oleh Suryabrata (1987) bahwa dalam pelaksanaan penilaian atau evaluasi yang menggunakan beberapa perangkat tes perlu dilakukan penyetaraan dari perangkat-perangkat tes yang dipakai, karena dengan penyetaraan perangkat tes dapat dijamin keadilan bagi peserta tes. Berdasarkan pendapat-pendapat di atas terlihat pentingnya proses penyetaraan dengan metode yang tepat bagi

perangkat tes atau paket tes yang lebih dari satu.

Melalui proses penyetaraan diperoleh tiga keuntungan pokok. Pertama, dapat digunakan perangkat tes yang berbeda terhadap kelompok yang berbeda sesuai dengan tingkat kemampuannya, sehingga skor yang diperoleh dapat dibandingkan. Selain itu peserta tes tidak merasa dirugikan atau diuntungkan karena mendapat tes yang lebih sukar atau lebih mudah. Kedua, bila terjadi kebocoran tes dari suatu perangkat tes tertentu dapat segera diganti dengan perangkat tes yang lain, yang sudah diketahui konstanta konversinya. Ketiga, fleksibilitas lingkungan dan waktu, artinya proses pengukuran dapat dilakukan pada tempat dan waktu yang berbeda jika kesetaraan paket tes tersebut sudah diketahui.

Butir-butir tes yang sudah disetarakan akan mempunyai satu skala ukurn. Adanya skala ukuran yang sama akan mempermudah pengontrolan mutu pendidikan. Oleh karena itu fungsi UAN sebagai pemantau mutu pendidikan nasional dapat dilaksanakan. Selain manfaat tersebut proses penyetaraan butir tes dengan metode penyetaraan yang tepat dapat digunakan untuk pengembangan penyusunan bank tes. Keberadaan bank tes sangat penting untuk penyusunan perangkat tes sesuai dengan kisi-kisi dan tujuan pengukuran. Sehingga tidak setiap proses pengukuran dilakukan penyusunan perangkat tes tersendiri. Keberadaan bank tes yang terkalibrasi dapat menekan pengeluaran biaya yang besar dalam setiap pembuatan tes.

Penyetaraan tes sangat dirasakan kegunaannya mengingat mutu pendidikan di Indonesia belum merata dan keadaan geografis Indonesia sebagai negara kepulauan yang cukup luas. Hal ini mengakibatkan pengukuran secara serentak dalam waktu yang sama tidak mudak untuk dilakukan, di samping itu untuk mengantisipasi pula hal-hal yang tidak diinginkan seperti kebocoran tes. Selain itu dengan adanya kebijaksanaan tentang otonomi pendidikan, maka pelaksanaan pendidikan menjadi hak dan wewenang daerah setempat. Oleh karena itu pengembangan bank tes di tiap-tiap daerah sangat diperlukan. Walaupun pelaksanaan pendidikan sudah menjadi wewenang daerah setempat, namun perintah pusat tetap berkewajiban untuk mengontrol kualitas pendidikan nasional. Misalnya dengan penetapan kurikulum dan kemampuan standar yang harus dimiliki oleh siswa pada jenjang pendidikan tertentu. Pengontrolan akan dapat dengan mudah dilaksanakan jika perangkat tes sebagai alat pengukuran atau bank tes yang ada di tiap-tiap daerah diketahui tingkat kesetaraannya. Tingkat kesetaraan perangkat tes yang berbeda akan dapat diketahui, jika dilakukan proses penyetaraan.

Secara rinci Lord (Hambleton & Swaminathan, 1985) mengungkapkan ada beberapa hal yang harus diperhatikan dalam penyetaraan tes, yaitu :

1. Perangkat tes yang mengukur sifat dan kemampuan yang berbeda tidak dapat disetarakan.
2. Skor mentah perangkat tes yang tidak sama reliabilitasnya tidak disetarakan.
3. Skor mentah perangkat tes yang memiliki tingkat kesukaran berbeda tidak dapat disetarakan.
4. Skor perangkat tes X dan Y tidak dapat disetarakan tanpa adanya bukti bahwa kedua perangkat tes paralel.

5. Skor-skor yang berasal dari dua perangkat tes yang berbeda materi tidak disetarakan.

Hubungan (*linking*) antartes dapat dikelompokkan menjadi 3, yakni penyetaraan (*equiting*), concordance, dan prediksi (*prediction*) (Kollen dan Brennan, 2004). Yang membedakan ketiga hubungan-hubungan ini adalah konstruk tes dan distribusinya. Jika Jika tes-tes tersebut secara statistik dan konseptual dapat saling menggantikan, maka hubungan dapat diketahui dengan penyetaraan (*equiting*), jika sama distribusinya (mengukur konstruk yang sama) dengan *concordance*, dan jika kondisi untuk penyetaraan dan *concordance* tidak terpenuhi, digunakan prediksi skor harapan. Fokus artikel ini diarahkan kepada persoalan penyetaraan tes (*test equating*). Di dalam proses penyetaraan paket tes, ada tiga hal yang perlu diperhatikan yaitu, rancangan penyetaraan yang digunakan, metode penyetaraan tes yang dipilih, dan arah penyetaraan tes. Berikut ini disajikan bahasan konseptual ketiga hal tersebut.

2. Rancangan Penyetaraan Tes

Salah satu hal yang diperhatikan dalam penyetaraan tes adalah menentukan rancangan penyetaraan. Ada tiga jenis rancangan yang dapat digunakan dalam penyetaraan tes, yaitu rancangan kelompok tunggal (RKT), rancangan kelompok ekuivalen (RKE), dan rancangan dengan butir jangkar (RBJ).

Dalam RKT (*single group design)* digunakan satu kelompok peserta yang merespons dua perangkat tes (X dan Y). Parameter butir dari kedua perangkat tes diestimasi secara terpisah dengan mengkalibrasi parameter kemampuan peserta (θ) atau parameter butir.

Dalam rancangan kedua, yaitu RKE (*equivalen group deesign)*, digunakan dua kelompok peserta ekuivalen (K_1 dan K_2) dan dua perangkat tes (X dan Y). Kelompok peserta K_1 mengerjakan perangkat tes X dan kelompok peserta K_2 mengerjakan perangkat tes Y. Mengingat kelompok K_1 dan K_2 adalah ekuivalen, maka kedua kelompok dianggap tunggal. Penentuan konstanta konversi berikutnya seperti Rancangan kelompok tunggal. Keuntungan rancangan ini dapat menghindari efek negatif yang disebabkan karena latihan dan kelelahan peserta tes, sedangkan kekurangannya ada kemungkinan bias yang disebabkan karena tidak mudah untuk membuat distribusi kemampuan dua kelompok peserta tes yang benar-benar ekuivalen.

Sedangkan pada rancangan ketiga, yaitu RBJ digunakan dua perangkat tes (X dan Y) dan dua kelompok peserta (K_1 dan K_2). Masing-masing perangkat tes ditambahkan item-item tes anchor Z, sehingga kedua perangkat tes menjadi (X+Z) dan (Y+Z). Kelompok peserta K_1 mengerjakan perangkat tes (X+Z) sedang kelompok peserta K_2 mengerjakan perangkat (Y+Z), sehingga item-item tes anchor Z dikerjakan oleh kedua kelompok peserta tes.

Pemilihan rancangan penyetaraan berhubungan dengan karakteristik tes yang akan disetarakan. Paket soal UAN yang berada tidak terdapat item tes anchor dan setiap peserta hanya mengerjakan satu perangkat soal. Dengan demikian penyetaraan tes dengan rancangan kelompok tunggal dan tes anchor tidak mungkin dilakukan.

3. Metode Penyetaraan

Ada beberapa metode penyetaraan tes yang dapat digunakan dan faktor-faktor yang mempengaruhi keakuratan metode penyetaraan tes. Dalam teori respos butir terdapat empat metode

penyetaraan tes, yaitu metode: regresi, rerata sigma, rerata da sigma tegar, dan kurva karakteristik (Anghoff, 1982; Lord, 1980). Keempat metode penyetaraan tersebut menggunakan prosedur yang berbeda-beda, sehingga ada kemungkinan konstanta konversi yang dihasilkan berbeda untuk penyetaraan paket tes yang sama. Penyetaraan paket tes yang sama walaupun dengan metode yang berbeda seharusnya didapatkan hasil yang sama pula.

Metode penyetaraan yang pertama adalah metode regresi. Penentuan konstanta konversi α dan β dengan menggunakan metode regresi dilakukan dengan memperhatikan respons peserta tes pada kedua perangkat tes X dan Y. Estimasi parameter butir dan parameter peserta memenuhi persamaan regresi linier, yaitu:

$$y = \alpha x + \beta + e$$

$$\alpha = r_{xy}\frac{S_y}{S_x}$$

$$\beta = \bar{y} - \alpha\bar{x}$$

Keterangan:

y : estimasi kemampuan atau estimasi parameter butir pada tes Y

x : estimasi kemampuan atau estimasi parameter butir pada tes x

r_{xy} : koefisien korelasi antara x dan y

$\bar{y}, \bar{x}$: rata-rata dari y dan x

S_x, S_y : standard deviasi dari x dan y

e : kesalahan dalam penaksiran garis regresi

Hambleton dan Swaminathan (1991) mengatakan bahwa kelemahan metode regresi tidak bersifat timbal balik (asimetris) sehingga kurang memadai untuk penentuan konstanta konversi. Lebih lanjut dinyatakan bahwa penyetaraan dua perangkat tes atau lebih memerlukan syarat invariansi dan timbal balik dari perangkat-perangkat tes yang disetarakan. Berdasarkan kenyataan tersebut metode Regresi tidak ikut dikomparasikan dalam penelitian ini karena dianggap kurang efisien dalam proses penyetaraan dan tidak memenuhi dalam penetapan kontanta penyetaraan.

Metode penyetaraan tes yang kedua adalah metode rerata sigma. Pada metode ini, penentuan konstanta konversi α dan β dengan menurut metode rerata dan sigma dilakukan dengan memperhatikan nilai estimasi parameter butir tes pada kedua perangkat tes yaitu b_x dan b_y. Metode rerata dan sigma bersifat timbal balik sehingga dengan cara yang sama hubungan dari y ke x dapat ditentukan. Menurut Hambleton dan Swaminathan (1985), hubungan antara estimasi parameter butir tes atau estimasi kemampuan peserta pada kedua perangkat tes yang akan disetarakan, memenuhi:

$$y = \alpha x + \beta$$

$$\bar{y} = \alpha\bar{x} + \beta$$

$$\alpha = \frac{S_y}{S_x}$$

$$\beta = y - \alpha x$$

Keterangan:

y : estimasi kemampuan atau estimasi parameter butir pada tes Y

x : estimasi kemampuan atau estimasi parameter butir pada tes X

$\bar{y}, \bar{x}$: rata-rata dari y dan x

S_x, S_y : standard deviasi dari x dan y

Metode penyetaraan tes yang ketiga disebut dengan metode rerata dan sigma tegar. Hambleton dan Swaminathan (1991), menyatakan bahwa dalam metode penyetaraan rerata

dan sigma tidak mempertimbangkan variasi estimasi parameter item. Linn et.al., (Hambleton dan Swaminathan, 1991) menyatakan bahwa metode penyetaraan rerata dan sigma tegar mempertimbangkan adanya variasi standard error estimasi parameter item.

Prosedur penyetaraan rerata dan sigma tegar dikembangkan oleh Linn, Levine, Hastings, dan Wardrop (dalam Hambleton dan Swaminathan, 1991). Langkah-langkah dalam penentuan konstanta konversi guna penyetaraan perangkat tes dengan menggunakan metode rerata dan sigma tegar adalah sebagai berikut.

a) Penentuan bobot parameter item (W_i), pada setiap pasangan (b_{xi},b_{yi}), yaitu:
$W_i = [maks\{v(x_i),v(y_i)\}]^{-1}$
$i = 1,2,3,...,k$
$v(x_i)$ dan $v(y_i)$ adalah Varians estimasi parameter tingkat kesulitan tes X dan Y.

b) Penentuan penskalaan bobot skala W_i dengan menggunakan rumus berikut.

$$w_i = \frac{w_i}{\sum_{j=1}^{k} w_j}$$

k : jumlah item anchor pada perangkat tes X dan Y

c) Penghitungan estimasi berbobot tes X dan Y, dengan menggunakan rumus:
$x'_i = w'_i x_i$
$y'_i = w'_i y_i$

d) Penentuan rerata dan simpangan baku dari estimasi berbobot tes X dan Y, yaitu x, y, S'x, S'y

e) Penentuan konstanta konversi α dan β dengan menggunakan rerata dan simpangan baku estimasi berbobot, dilakukan deengan mensubstitusikan rerata dan simpangan baku bobot estimasi pada persamaan penyamaan skala.

Menurut Stocking dan Lord (Hambleton, 1985) dalam metode penyetaraan rerata dan sigma, proses penentuan konstanta konversi tidak memperhatikan kemungkinan skor kelompok ekstrim, sedangkan metode penyetaraan rerata dan sigma tegar dapat diperbaiki dengan jalan memperhatikan skor kelompok ekstrim.

Menurut Stocking dan Lord (Hambleton dan Swaminathan, 1985), langkah-langkah penentuan konstanta konversi pada kelompok ekstrim pada dasarnya seperti langkah a) sampai e), dilanjutkan langkah-langkah berikut:

f) Dengan menggunakan nilai α dan β yangsudah ditentukan jarak pasangan (x_i, y_i) terhadap garis penyetaraan, adalah:

$$d_i = \left| \frac{(y_i - \alpha x_i - \beta_i)}{\sqrt{\alpha^2 + \beta^2)}} \right|$$

g) Jika M adalah median dari di, kemudian dilakukan perhitungan bobot Tukey dengan fungsi:

$$T = \left\{ \left[1 - \left(\frac{d_i}{6M} \right)^2 \right]^2 \text{ untuk } d_i < 6M \right.$$

h) Pembobotan tiap-tiap pasangan (x_i,y_i) dengan rumus:

$$u_i = w'iT \left\{ \sum_{i=1}^{n} w'iT_i \right\}$$

i) Ulangi langkah c) dengan menggunakan ui sebagai pengganti W'_i, kemudian ditentukan α dan β seperti langkah e)

j) Ulangi langkah f) sampai i) sampai didapatkan hasil α dan β lebih kecil dari α dan β yang sudah ditentukan.

Sedangkan metode keempat yang dapat digunakan dalam penyetaraan

tes adalah metode kurva karakteristik. Penentuan konstanta konversi α dan β dengan metode kurva karakteristik, dilakukan dengan memperhatikan nilai estimasi parameter butir tes kedua perangkat soal yaitu x dan y. Metode penyetaraan rerata dan sigma serta metode rerata dan sigma tegar dalam penentuan konstanta konversi hanya memperhitungkan hubungan yang ada antara parameter-parameter kesukaran butir pada perangkat tes yang satu terhadap perangkat tes yang lain. Hubungan antara parameter-parameter daya beda pada kedua perangkat tes belum dipertimbangkan.

Haebara (1980), menyatakan bahwa metode kurva karakteristik mempertimbangkan informasi dari parameter daya beda butir dan tingkat kesukaran butir dalam penentuan konstanta konversi. Oleh karena itu, dalam Metode penyetaraan kurva karakteristik diperhatikan hubungan antara parameter-parameter kesukaran daya beda dan hubungan antara parameter kesukaran butir tes yang akan disetarakan. Selain itu juga dalam metode kurva karakteristik diperhatikan skor asli (*true score)* peserta tes pada kedua perangkat tes.

True Score (t_{xa}) dari peserta tes dengan kemampuan θ_a yang merespon k item dalam perangkat tes X dan tes Y adalah:

$$\tau_{xa} = \sum_{i=1}^{k} p(\theta_a, b_{xi}, a_{xi}, c_{xi})$$

$$\tau_{ya} = \sum_{i=1}^{k} p(\theta_a, b_{yi}, a_{yi}, c_{yi})$$

Setiap item pada perangkat tes X dan Y memenuhi persamaan:

$$b_{yi} = \alpha b_{xi} + \beta$$

$$a_{yi} = \frac{a_{xi}}{\alpha} \quad \text{atau}$$

$$\alpha = \frac{a_{xi}}{a_{yi}}$$

$$c_{yi} = c_{xi}$$

$$\beta = b_{yi} - \alpha b_{xi}$$

Konstanta α dab β dipilih sedemikian sehingga fungsi F seperti tertera di bawah ini mencapai nilai minimal.

$$F = \frac{1}{N} \sum_{a=1}^{N} (\tau_{xa} - \tau_{ya})$$

Keterangan:

F : fungsi dari α dan β, yang menunjukkan ketidaksesuaian antara τ_{xa} dan τ_{ya}

N : jumlah peserta tes

τ_{xa} : true score peserta tes pada kemampuan a pada perangkat tes X

τ_{ya} : true score peserta tes pada kemampuan a pada perangkat tes Y

Untuk menentukan nilai minimal fungsi F, digunakan pendekatan numerik *(Golden Section"* (Chapra dan Canale, 1996; Susila, 1994).

4. Bentuk-Bentuk Penyetaraan Tes

Kolen & Brennan (1994), memilah penyetaraan tes menjadi dua, yaitu penyetaraan tes vertikal dan penyetaraan tes horisontal. Metode penyetaraan tes vertikal adalah penyetaraan tes yang digunakan antarlevel yang berbeda. Misalnya tes untuk mengukur kemampuan matematika kelas I, II, dan III. Untuk kepentingan penyetaraan tes vertikal, tes dapat dirancang tanpa atau dengan butir jangkar. Dalam pelaksanaannya, tes yang dikembangkan tanpa

menggunakan butir jangkar diujikan untuk kelas I, II, dan III dengan waktu pengerjaan tes diperpanjang secara proporsional. Apabila satu paket tes diujikan selama 1 jam, maka apabila tiga paket tes diujikan sekaligus waktu dilipatkan menjadi 3 jam.

Namun, dalam hal demikian harus diperhatikan kemungkinan kelelahan peserta tes yang mengakibatkan respon tidak valid. Model penyetaraan tes vertikal yang kedua adalah dengan menggunakan butir jangkar. Tes dikembangkan menjadi tiga paket, paket pertama untuk kelas I, paket kedua untuk kelas II, dan paket ketiga untuk kelas III. Setiap paket tes dimasukkan unsur butir jangkarnya. Jumlah butir jangkar yang digunakan minimal 20% dari jumlah butir tes (Skaggs & Lissitz, 1986). Penyetaraan tes vertikal seringkali digunakan di Amerika pada baterai tes prestasi jenjang sekolah dasar. Penyetaraan tes vertikal tidak ditujukan untuk menyesuaikan antarpaket tes sehingga skor tes bisa saling menggantikan, karena tes ini mengukur kemampuan pada level dan materi yang berbeda. Tujuan tes ini adalah untuk mengetahui tingkat perkembangan kemampuan anak dalam menguasai materi.

Sedangkan penyetaraan tes horisontal adalah penyetaraan tes dimana terdapat dua paket tes atau yang dikembangkan berdasarkan isi dan item tes yang sama, namun lazimnya setiap paket tes memiliki perbedaan tingkat kesulitan. Setiap kelompok peserta tes mengerjakan paket tes berbeda yang memiliki butir jangkar. Skor tes yang diperoleh peserta tes pada setiap kelompok selajuntya dapat saling menggantikan dengan metode penyetaraan yang ada.

Kolen & Brennan (dalam Chong Ho Yu dan Sharon E. Osborn Popp, 2005), mengemukakan ada empat aspek kesetaraan yang harus diperhatikan dalam penyetaran tes. Keempat aspek itu adalah :

1. Interferensi
   Seberapa jauh skor dari kedua tes dapat digunakan untuk mengukur tujuan yang sama. Misalnya mengukur prestasi akuntansi, mengukur kemampuan berhitung.
2. Konstruk
   Seberapa jauh kedua paket tes mengukur konstruk yang sama.
3. Populasi
   Seberapa jauh populasi yang digunakan adalah homogen atau sama. Selain itu faktor-faktor kualitas dan kuantitas yang berhubungan dengan sistem pembelajaran harus disetarakan. Artinya sekolah yang memiliki siswa dengan latar belakang sosial dan ekonomi jauh di bawah, fasilitas sarana prasarana sekolah serba kekurangan, dan guru yang seadanya tidak tepat bila dibandingkan dengan keadaan yang tidak setara.
4. Karakteristik atau kondisi pengukuran
   Seberapa jauh kesamaan kondisi pengukuran dilakukan untuk kedua paket tes, baik dari sisi panjang tes, bentuk tes, adminstrasi tes, waktu tes, tipe item, dan prosedur tes.

5. Prosedur Penyetaraan Tes

a. Uji Prasyarat

Untuk melakukan uji kesetaraan tes sering dikatakan pula sebagai uji kepararelan tes, sehingga paket tes yang tidak paralel diperlukan penyetaraan tes agar tidak merugikan kelompok tertentu dalam hal menentukan keputus-

an. Selanjutnya dilakukan uji post hoc untuk mendapatkan pasangan paket tes mana yang setara dan mana yang tidak setara. Uji post hoc dapat dilakukan dengan uji Scheffe, Tukey, Bonferroni, LSD (*least significance differences*), atau metode lain. Syarat untuk menguji kesetaraan tes adalah distribusi normal dan varians skor kelompok homogen. Uji normalitas distribusi dilakukan dengan Kolmogorov Smirnov test, sedangkan untuk menguji homogenitas varians digunakan Levene test (Sudjana, 1983).

Ada tiga jenis rancangan yang dapat digunakan dalam penyetaraan tes, yaitu rancangan kelompok tunggal (RKT), rancangan kelompok ekuivalen (RKE), dan rancangan dengan butir jangkar (RBJ). Apabila penyetaraan tes menggunakan RKT, maka satu kelompok yang terdiri dari 15 orang tersebut mengerjakan ketiga paket tes UAN (A, B, C). Pada RKE, ada tiga kelompok dengan setiap kelompok terdiri dari 15 orang. Kelompok pertama mengerjakan paket tes A, kelompok kedua mengerjakan paket tes B, dan kelompok ketiga mengerjakan paket tes C. Sedangkan pada RBJ, ketiga paket tes memiliki butir yang sama yang disebut dengan butir jangkar. Dalam hal RBJ digunakan, maka harus ada kelompok sejumlah paket tes. Kelompok pertama mengerjakan paket tes A, kelompok kedua mengerjakan paket tes B, dan kelompok ketiga mengerjakan pakeet tes C. RBJ ini disebut juga dengan nama *common item nonequivalent groups design.*

Misalkan ada tiga paket tes UAN ekonomi akuntansi SMU (sebut A, B, C) yang akan diuji kesetaraannya dengan RKE. Pemilihan RKE karena UAN dirancang dengan untuk kelompok ekuivalen dan paket tes UAN tidak menggunakan butir jangkar dan Ketiga paket tes UAN tersebut masing-masing diikuti oleh lima belas orang peserta tes. Uji kesetaraan paket soal dilakukan dengan uji beda rerata dengan analisis varians dan uji pasang (*post hoc test*) dengan analisis LSD. Nilai ketiga kelompok pada mata pelajaran ekonomi akuntansi SMU tersebut adalah sebagai berikut.

Subjek / Paket	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
Paket A	9	9	6	8	8	7	6	5	3	6	7	6	5	7	8
Paket B	6	6	5	6	9	5	4	3	5	5	6	6	6	5	5
Paket C	2	3	6	2	3	3	2	2	3	2	2	4	6	5	2

Berdasarkan skor peserta tes itu selanjutnya diuji beda mean untuk mengetahui kesetaraan nilai setiap kelompok. Berdasarkan analisis anova satu jalur dan uji beda varians dengan SPSS 11.5 diperoleh hasil sebagai berikut.

a. Hasil uji beda rerata total

ANOVA

VAR00001

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	96.844	2	48.422	22.398	.000
Within Groups	90.800	42	2.162		
Total	187.644	44			

b. Hasil uji post hoc dengan LSD

Multiple Comparisons

Dependent Variable: VAR00001
LSD

(I) VAR00002	(J) VAR00002	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	1.2000*	.53689	.031	.1165	2.2835
	3.00	3.5333*	.53689	.000	2.4498	4.6168
2.00	1.00	-1.2000*	.53689	.031	-2.2835	-.1165
	3.00	2.3333*	.53689	.000	1.2498	3.4168
3.00	1.00	-3.5333*	.53689	.000	-4.6168	-2.4498
	2.00	-2.3333*	.53689	.000	-3.4168	-1.2498

*. The mean difference is significant at the .05 level.

c. Hasil uji homogenitas varians

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
SKOR_AI	Equal variances assumed	1.335	.259	2.225	28	.034	1.2000	.53925	.09339	2.30661
	Equal variances not assumed			2.225	26.677	.035	1.2000	.53925	.09292	2.30708

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
SKOR_AI	Equal variances assumed	.233	.633	6.282	28	.000	3.5333	.56512	2.37573	4.69094
	Equal variances not assumed			6.282	27.645	.000	3.5333	.56512	2.37508	4.69161

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
SKOR_B	Equal variances assumed	.504	.484	4.624	28	.000	2.3333	.50458	1.29974	3.36692
	Equal variances not assumed			4.624	27.692	.000	2.3333	.50458	1.29916	3.36751

d. Hasil uji normalitas skor tiga kelompok

One-Sample Kolmogorov-Smirnov Test

		VAR001	VAR002	VAR003
N		15	15	15
Normal Parameters[a,b]	Mean	6.8667	5.4667	3.1333
	Std. Deviation	1.63299	1.30201	1.45733
Most Extreme Differences	Absolute	.142	.274	.270
	Positive	.125	.274	.270
	Negative	-.142	-.227	-.218
Kolmogorov-Smirnov Z		.548	1.063	1.045
Asymp. Sig. (2-tailed)		.925	.209	.225

a. Test distribution is Normal.
b. Calculated from data.

Berdasarkan uji homogenitas varians dengan uji Levene diperoleh kesimpulan bahwa varians antarkelompok homogen, yang dibuktikan dengan semua nilai signifikansi F untuk lebih besar dari 0.05. Berdasarkan uji normalitas distribusi dengan Kolmogorov Sminorv test dapat disimpulkan bahwa distribusi skor ketiga kelompok adalah normal, yang dibuktikan dengan nilai signifikansi di atas 0,05.

Hasil uji beda rerata ketiga paket tes didapat nilai F = 22,398 dengan tingkat signifikansi 0,000 berarti rerata ketiga paket tes dari ketiga populasi berbeda secara signifikan. Selanjutnya berdasarkan uji post hoc dengan LSD dapat disimpulkan ketiga paket tes tidak setara, karena semua nilai signifikansi uji beda antarkelompok lebih kecil dari 0,05. Berdasarkan uji prasyarat tersebut dapat disimpulkan bahwa ketiga paket tes tersebut tidak setara sehingga perlu dilakukan penyetaraan tes.

b. Estimasi Parameter Butir dan Kemampuan

Setelah diketahui bahwa ketiga paket tes UAN tidak setera maka tahap berikutnya dalam penyeteraan tes adalah mengestimasi parameter butir. Estimasi butir tes dapat dilakukan dengan program BILOG, MULTILOG, atau EXCEL. Estimasi ini dilakukan untuk menentukan nilai parameter daya beda (a), tingkat kesulitan (b), terkaan semu (c), dan estimasi kemampuan peserta tes (θ). Estimasi ini dilakukan untuk menentukan model parameter logistik (PL) yang akan digunakan (1 PL, 2 PL, atau 3 PL) yang cocok dengan data respon peserta tes. Analisis dengan BILOG akan menghasilkan tiga output, yaitu estimasi butir berdasarkan teori tes klasik, estimasi butir dengan teori respon butir (*item response theory*), dan tahap estimasi kemampuan peserta tes (θ). File perintah dan data mentah untuk menjalankan program bilog dapat ditulis dengan program underdos, editplus, atau notepad. Nama file perintah menggunakan ekstensi "blg" misalnya S10.blg. Pada baris tertentu dituliskan model parameter

logistik yang ingin dipakai dengan perintah >SCORE RSC=3; untuk 3 PL, diisi angka 2 untuk 2 PL, dan diisi 1 untuk 1 PL. Berikut isi file perintah S10.blg tersebut.

```
>COMMENTS
KARAKTERISTIK MODEL 3P
SUKIRNO DS'
>GLOBAL
DFNAME='c:\bilog\DBIL3b\S2711.DA
T',KFNAME='c:\bilog\DBIL3b\S2711.
DAT',
OENAME='c:\bilog\DBIL3b\S2711.DA
T',NPARM=3,OMITS,SAVE;
>SAVE
PARM='c:\bilog\DBIL3b\S2711.PAR';
>LENGTH NITEMS=10;
>INPUT
NTOT=10,NALT=4,NIDC=5,SAM=15;
(1X,5A1,T6,10A1)
>TEST TNAME=AKTMAN;
>CALIB FLOAT;
>SCORE RSC=3;
```

Kemudian, file data yang ditulis juga dengan program underdos, editplus, atau notepad disimpan dengan ekstensi "dat", misalnya S10.dat. Isi file S10.dat adalah sebagai berikut.

```
AK 1111111111
OK 9999999999
 1 0000100000
 2 1110010000
 3 1010111100
 4 0110000000
 5 0100101000
 6 0000011001
 7 0110000000
 8 1010000000
 9 1100010000
10 0010010000
11 0010010000
12 1110001000
13 1010111010
14 1101010001
15 1100000000
```

Karena ada tiga kelompok dan tiga paket tes, maka file perintah dan data dibuat sejumlah kelompok dan data respon peserta tesnya. Berdasarkan output BILOG pertama dapat dianalisis butir-butir tes mana yang harus diperbaiki atau dihapus dari analisis karena kualitas butir tes tidak memenuhi standar teori klasik (daya beda, tingkat kesulitan, distraktor). Pada bagian kedua output BILOG disajikan estimasi parameter butir (daya beda, tingkat kesulitan, dan terkaan semu) sesuai dengan model logistik yang akan digunakan. Pada bagian akhir output disajikan estimasi kemampuan peserta tes. Berdasarkan tiga output estimasi itulah selanjutnya dapat diestimasi persamaan penyetaraan tes.

Untuk menjalankan file program dapat menggunakan BILOG underdos atau *underwindow*. Perintah yang ditulis dari *prompt* untuk menganalisis data dengan BILOG underdos adalah sebagai berikut.

C:\bilog>**bilog s10.blg** (enter)

Sedangkan apabila menggunakan BILOG underwindow maka perintah yang dipilih adalah sebagai berikut.

1. File Open (pilih file s10.blg)
2. File Run

Selanjutnya akan diperoleh hasil analisis BILOG yang teridiri dari tiga output, yang terdiri dari S10.PH1 (berisi estimasi butir berdasar teori klasik), S10.PH2 (berisi estimasi butir dengan IRT), dan S10.PH3 (berisi estimasi kemampuan peserta tes) yang dapat digunakan untuk menentukan model parameter logistik (PL) yang akan digunakan (1 PL, 2 PL, atau 3 PL) yang cocok dengan data respon peserta tes.

Setelah di perintah program ditulis-kan model logistik yang akan diguna-

kan, kemudian dijalankan program tersebut, maka akan diperoleh nilai χ^2 hitung dan skor probabilitas yang terletak dalam file s10.PH2 pada bagian tengah output. Apabila nilai χ^2 hitung untuks etiap butir tes lebih besar atau sama dengan dari nilai χ^2 tabel pada dk =1 atau skor probabilitas lebih kecil atau sama dengan 0,05, berarti model tersebut tidak cocok dengan model yang dipilih. Kegiatan ini dilakukan untuk semua paket tes (A, B, C).

c. Estimasi Persamaan Penyetaraan

Setelah diketahui model logistic yang cocok, selanjutnya dapat ditemukan butir-butir yang cocok. Butir-butir tes tersebut kemudian digunakan kembali untuk menentukan skor peserta tes (berapa jumlah respon yang benar). Skor peserta itulah yang digunakan untuk menentukan nilai penyetaraan sesuai dengan metode penyetaraan yang dipakai.

Dalam hal penyetaraan paket tes A dan paket tes B, maka skor paket tes A dianggap sebagai X dan skor paket tes B, sebagai Y atau sebaliknya. Setelah proses penyetaraan tes dilakukan akan diperoleh nilai parameter konversi (α dan β) yang diperoleh dari masing-masing metode penyetaraan yang digunakan. Mengingat kompleks dan banyaknya prosedur yang harus dilalui untuk menyusun persamaan penyetaraan tes, maka dalam tulisan ini hanya dibatasi pada penyetaraan tes menggunakan metode regresi dengan rancangan kelompok ekuivalen.

Misalkan nilai parameter konversi A – B, A - C, dan B – C adalah sebagai berikut.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3.596	1.745		2.061	.060
	PAKET_B	.562	.311	.448	1.806	.094

a. Dependent Variable: PAKET_A

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	7.650	1.025		7.465	.000
	PAKET_C	-.314	.298	-.280	-1.052	.312

a. Dependent Variable: PAKET_A

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	5.143	.845		6.083	.000
	PAKET_C	.103	.246	.115	.419	.682

a. Dependent Variable: PAKET_B

Selanjutnya, untuk membuat penyetaraan paket tes A – C dengan ketiga model digunakan persamaan sebagai berikut.

Penyetaraan A ke B atau B ke A : A = 3,596 + 0,562 B

Penyetaraan A ke C atau C ke A : A = 7,650 – 0,341 C

Penyetaraan B ke C atau C ke B : B = 5,143 + 0,103 C

Penggunaan hasil penyetaraan itu misalnya, orang ke satu pada paket tes A memperoleh skor 9, apabila ditransfer nilainya ke paket tes B = 9,62 dengan perhitungan sebagai berikut.

9 = 3,596 + 0,562 B

B = (9 - 3,596) / 0,562

B = 9,62

Demikianlah prosedur penyetaraan tes yang dilakukan dengan menggunakan metode regresi. Prosedur penyetaraan tes digunakan agar tidak ada peserta UAN yang dirugikan dan nilai yang diberikan kepada setiap peserta tes UAN menjadi adil meskipun setiap peserta tes mengerjakan paket tes berbeda dan tingkat kesulitan yang tidak sama.

C. Simpulan dan Saran

1. Simpulan

Berdasarkan deskripsi konseptual tentang latar belakang pentingya penyetaraan tes UAN dan cara penyetaraan tes di atas, dapat disimpulkan tiga hal sebagai berikut.

(1) Penyetaraan tes UAN diperlukan karena ada tiga keuntungan yang diperoleh.
   (a) Dapat digunakan perangkat tes yang berbeda terhadap kelompok yang berbeda sesuai dengan tingkat kemampuannya, sehingga skor yang diperoleh dapat dibandingkan. Selain itu peserta tes tidak merasa dirugikan atau diuntungkan karena mendapat tes yang lebih sukar atau lebih mudah.
   (b) Bila terjadi kebocoran tes dari suatu perangkat tes tertentu dapat segera diganti dengan perangkat tes yang lain, yang sudah diketahui konstanta konversinya.
   (c) Fleksibilitas lingkungan dan waktu, artinya proses pengukuran dapat dilakukan pada tempat dan waktu yang berbeda jika kesetaraan paket tes tersebut sudah diketahui.

(2) Di dalam proses penyetaraan paket tes, ada tiga hal yang perlu diperhatikan yaitu, rancangan penyetaraan yang digunakan, metode penyetaraan tes yang dipilih, dan arah penyetaraan tes.

(3) Prosedur penyetaraan tes digunakan agar tidak ada peserta UAN yang dirugikan dan nilai yang diberikan kepada setiap peserta tes UAN menjadi adil meskipun setiap peserta tes mengerjakan paket tes berbeda dan tingkat kesulitan yang tidak sama.

2. Saran

(1) Informasi tentang paket tes dan penyetaraan tes yang digunakan perlu dipublikasi agar masyarakat dapat mengevaluasi secara proporsional penyelenggaraan UAN.

(2) Prosedur penyetaraan tes sangat penting dipahami oleh para guru dan dosen, oleh karena itu pelatihan tentang penyetaraan tes untuk para guru dan dosen mendesak dilakukan.

(3) Metode penyetaraan tes yang baik terdiri dari berbagai tahapan yang rumit, oleh karena itu perlu diajarkan prosedur penyetaraan tes yang lain yang lebih sederhana dan mudah dipelajari dan dipraktikkan oleh para guru dan dosen.

Daftar Pustaka

Angoff WH. 1982. Uses of Difficulty and Discrimination Indices for Detecting Item Bias In RA Berk. *Handbook of Methods for Detecting Item Bias*. Baltimore: Johns Hopkins University Press.

Chong Ho Yu dan Sharon E. Osborn Popp, 2005. Test Equating by Common Items and Common Subjects: Concepts and Applications. *Practical Assessment, Research & Evaluation.* Volume 10 Number 4, May 2005.

Dananwijaya. 2000. Mengapa UAN harus Dihapus. *Kompas*. hal. 9 Tanggal 19 Juni 2000.

Depdikbud. 1998. *Pedoman Kegiatan Penulisan Usul Soal UAN SD, MI, SLTP/MTS, SMU/MA, SMK Tahun Pelajaran 1998/1999*. Semarang: Kanwil Depdikbud.

Ebel, R L & Frisbie, D. A. 1986. *Essentials* of *Educational Measurement*. Englewood Cliffs, NJ: Prentice Hall Inc.

Ebel R.L. 1979. *Essentials of Educational Measurement*. 3rd. Edition Englewood Cliffs, New Jersey: Prentice Hall, Inc.

Erna Miyatun & Djemari Mardapi. 2000. Komparasi Metode Penyetaraan Tes Menurut Teori Respon Butir. *Jurnal Penelitian dan Evaluasi*. Nomor 3 Tahun II, 2000.

Feldt , L. S & Chorter, R. A. 2003. Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Methods, 8 (1), 102*109.

Gronlund, N. E. 1990. *Measurement and Evaluation in Teaching*. New York: Macmillan Publishing Co. Inc.

Hambleton RK. & Linda L. Cook. 1997. Latent Trait Models and Their Use in the Analysis of Education Test Data. *Journal of Educational Measurement*. 14. hal. 75 – 96.

Hambleton, R K. 1989. Principles and Selected Applications of Item Response Theory. Dalam RL. Linn. *Education Measurement. Hal. 147-200.* New York: Macmillan.

Hambleton, R. K & Jones, R. W. 1994. Item Parameter Estimation Errors and Their Influence on Test Information Functions. *Applied Measurement in Education*, 7(3), pp. 171-186.

Hambleton RK, Swaminathan H, & Rogers HJ. 1991. *Fundamentals of Item Response Theory*. Newbury Park : Sage Publications Inc.

Hambleton,R .K., & Swaminathan. H., 1985. Item Response Theory: Principles and Applications. Boston: Kluwer: Nijhoff Publishing.

_____, 1989. Applications of Item Response Theory to Practical Testing Problems. New Jersey : Lawrence Erlbaum Associates Publisher.

Hambleton R, .K., Swaminathan,H ., dan Rogers, H.I. 1991. Fundamental Item Response Theory. London: Sage Publications, Inc.

Hayat, Bahrul & Surya Pranata. 1995. *Analisis dan Kaliberasi Soal UAN SMP Tahun Pelajaran 1993/1994*. Jakarta : Puslitbang Sisjian Depdikbud.

HJ Rogers dalam Keeves. 1999. Guessing in Multiple Choice Test. *Advances in Measurement in Educational Research and Assessment*, 1999, Second Edition, Amsterdam : Elsevier Science Ltd. pp. 235 – 243.

Kaplan RM. & Sacuzzo DP. 1982. *Psychological Testing Principles*, Applications, and Issues. Monetery California : Brooks / Cole.

Kolen, Michael J. & Robert L. Brennan. 1995. *Test Equating*. New York : Springer Verlag New York Inc.

Linn, R L & Gronlund, N. F. 2000. *Measurement and Assessment in*

Teaching. Upper Saddle River N.J : Prentice-Hall Inc.

Mardapi, D. 1997. Ragam Bentuk Evaluasi. *Makalah Semiloka Evaluasi Sistem Penilaian dan Pengukuran Hasil Belajar Mahasiswa UGM*, di Universitas Gadjah Mada.

Mardapi Djemari. 1998. Analisis Butir dengan Teori Klasik dan Teori Respon Butir, *Jurnal Kependidikan*. Edisi Khusus Dies Tahun XXVIII, 1998.

Mislevy, Robert J dan R. Darrel Bock. 1990. *Bilog 3. Item Analysis and Test Scoring with Binary Logistic Models. Second Edtion*. Mooresville : Scientific Software Inc.

Nonny Swediati. 1997. Equating Tests under the Generalized Partial Credit Model. *Disertasi*. Tidak Dipublikasikan.University of Massachusetts at Amherst.

Holland PW & Rumib DB. 1982. Stroud TWF : Discussion of a Test of the Adequacy of Linear Score Equating Models. *Test Equating*. Hal. 137 – 138. New York : Academic Press Inc.

Skaggs G. & Lissitz RW. 1986. IRT Equating: *Relevant Issues and a Review of Recent Research. Review of Educational Research*. 56. hal. 495 – 529.

Stuart Luppescu. 2005. Virtual Equating. *Rasch Measurement*. Vol. 19 No. 3 Winter 2005

Sudjana. 1983. *Dasar-Dasar Statistika*. Bandung : Tarsito.

Suryabrata, S. 1984. *Pembimbing ke Psikodiagnostik (edisi ke 2). Yogyakarta:* Sarasin

_____. 1987. *Pengembangan Tes Hasil Belajar*. Jakarta: CV. Raja

Supriyoko. 2000. Jangan Jadikan Guru Sebagai Pembohong. *Republika*. hal. 11 Tanggal 12 Juni 2000.

Wright BD. & Stone M. 1797. *Best Test Design*. Chicago : Mesa Press.