PENGARUH JUMLAH BUTIR *ANCHOR* TERHADAP HASIL PENYETARAAN TES BERDASARKAN TEORI RESPON BUTIR

Syahrul, Mansyur, dan Rosdiyanah
Fakultas Teknik Universitas Negeri Makassar
email: syahrulab@yahoo.co.id

Abstrak

Penelitian ini bertujuan untuk mengetahui hasil penyetaraan tes berdasarkan perbedaan jumlah butir *anchor* dan prosedur mendapatkan penyetaraan tes berdasarkan *equateIRT*. Jenis penelitian ini adalah exploratif yaitu mengungkap kesetaraan skor tes berdasarkan teori respon butir. Instrumen yang digunakan dalam penelitian ini adalah enam paket soal Fisika. Penelitian ini dilaksanakan pada enam SMA di Kabupaten Gowa. Jumlah subjek penelitian sebanyak 1420 siswa. Desain penyetaraan memilih *Common-Item Nonequivalent Group*, estimasi parameter menggunakan model logistik dua parameter (2PL), dan penyetaraan tes dengan menggunakan *equateIRT*. Hasil penelitian menunjukkan bahwa koefisien penyetaraan α dan β yang dihasilkan oleh paket soal dengan 16 butir soal *anchor* (40%) lebih mendekati α = 1 dan β = 0. *Standard error* yang dihasilkan oleh paket soal dengan 16 butir *anchor* lebih kecil dibandingkan dengan paket soal dengan 10 dan 12 butir *anchor*. Hal ini berarti bahwa paket soal dengan jumlah butir *anchor* yang lebih banyak menghasilkan penyetaraan yang lebih akurat.
Kata kunci: *butir anchor, penyetaraan tes, teori respon butir*

THE INFLUENCE OF ANCHOR ITEM TOWARD THE EQUATING TESTS OUTCOMES BASED ON ITEM RESPONSE THEORY

Abstract

This study was aimed at finding out the equating test outcome based on the differences of numbers of anchor items and procedures to obtain equivalency tests based on equate IRT. This was an explorative research on the equality of test scores based on the item response theory. The instrument used in this study included six test packages of Physics. The research was conducted at six senior high schools in Gowa regency. The subjects were 1,420 students. The equating design used was Common-Item non-equivalent Group, while the parameter estimation used was the two-parameter logistic model (2PL), and test equating used was equate-IRT. The research results show that the equalization coefficients α and β are generated by a package of 16 items about the anchor (40%) approximates α = 1 and β = 0. The standard error generated bythe package of 16 items about the anchor is smaller than the package about with 10 and 12 point anchor. This means that a package about the amount of grains that produces more anchors produces more accurate equalization.
Keywords: *anchor item, test equating, item response theory*

PENDAHULUAN

Pengukuran hasil belajar di sekolah terutama hasil belajar kognitif dilakukan dengan menggunakan alat ukur yang dinamakan tes. Alat ukur yang baik memberikan hasil yang konstan bila digunakan berulang-ulang, jika kemampuan yang diukur tidak berubah. Ketepatan alat ukur dapat dilihat dari konstruk alat ukur, yaitu mengukur seperti yang direncanakan. Pengukuran yang tepat dapat memberikan informasi yang akurat mengenai penguasaan seseorang atau sekelompok orang terhadap materi yang dipelajari dan informasi itu berguna untuk membuat sebuah keputusan pendidikan.

Pengukuran modern bertujuan untuk menghilangkan kelemahan pada pengukuran klasik. Tujuan utama pengukuran modern adalah melepaskan keterpisahan di antara butir uji tes dengan peserta uji tes. Dengan pengukuran modern ini, ciri butir akan tetap sama, tidak menjadi soal peserta yang menempuhnya. Demikian pula, ciri peserta akan tetap sama, tidak menjadi soal butir yang ditanggapainya.

Salah satu jenis pengukuran modern yang cukup terkenal adalah teori respon butir yang dikembangkan atas dasar dua postulat yaitu: (1) performansi subjek pada suatu butir dapat diprediksikan oleh seperangkat faktor yang disebut *latent trait* atau kemampuan dan (2) hubungan performansi subjek pada suatu butir dan perangkat kemampuan laten yang mendasarinya digambarkan oleh fungsi naik monoton yang disebut *Item Charascteristic Curve (ICC)*. Selain itu, asumsi-asumsi yang melandasi teori respon butir adalah unidimensi, independensi lokal, dan fungsi karakteristik butir atau kurva karakteristik butir (Hambleton, Swaminathan, & Rogers, 1991).

Penilaian hasil belajar peserta didik pada dasarnya merupakan bagian integral dari proses pembelajaran (Taruno, 2014). Seringkali dijumpai di sekolah, peserta tes harus diukur dengan tes yang berbeda, meskipun tes-tes itu belum tentu ekivalen dan diharapkan dapat diukur sifat serta tuntutan pencapaian hasil yang dapat dibandingkan. Hal ini terjadi, misalnya pada situasi sekolah yang memiliki beberapa kelas paralel yang diajar oleh guru yang berbeda atau ketika guru memberikan ulangan susulan kepada siswa-siswa yang absen pada saat ulangan dilaksanakan. Meskipun sampai taraf tertentu, kesetaraan beberapa tes dapat diupayakan pada saat menyusun tes-tes itu sendiri. Akan tetapi, umumnya variasi taraf kesukaran antartes tetap terjadi.

Kenyataan menunjukkan bahwa masih banyak guru belum mengetahui prosedur pembuatan tes yang baik (Syahrul, 2014). Kebanyakan tes disusun dalam jangka waktu yang sangat singkat bahkan ada juga yang mengadopsi langsung butir-butir tes yang telah tersedia dalam buku panduan belajar sehingga perangkat tes yang digunakan oleh guru tidak dapat mengukur yang sebenarnya akan diukur. Seorang guru harus mengetahui dasar-dasar penyusunan tes prestasi belajar yang baik agar dapat memperoleh hasil ukur yang akurat (valid) dan dapat dipercaya (reliabel).

Dalam penyelenggaraan evaluasi hasil belajar, guru terkadang kesulitan untuk menyelenggarakan tes, misalnya tes formatif, sumatif, dan tes kenaikan kelas untuk kelas pararel yang cukup banyak. Para guru menggunakan satu perangkat tes saja sehingga tidak menutup kemungkinan siswa saling mencontek atau terjadi kebocoran soal. Kondisi tersebut berakibat pada pemberian nilai/skor terhadap hasil belajar siswa tidak mencerminkan kemampuan sebenarnya. Hal ini seperti dikemukakan oleh Rosana dan Sukardiyono (2015) bahwa

ada kalanya skor peserta didik tidak sesuai dengan kemampuannya yang sebenarnya. Penyebabnya dapat diakibatkan oleh permasalahan yang muncul dari peserta didik dan dapat juga diakibatkan oleh kualitas butir tes yang diberikan pada mereka sehingga nilai yang diberikan kepada siswa atau peserta tes lebih tinggi atau lebih rendah dari kemampuan prestasi sebenarnya.

Untuk menghindari situasi demikian, para guru juga membuat lebih dari satu perangkat tes (tes paralel) yang memiliki kisi-kisi yang sama dan untuk mengukur tingkat kemampuan yang sama. Akan tetapi, para guru belum memiliki kemampuan untuk melakukan analisis butir soal, terutama untuk menentukan perangkat-perangkat tes tersebut memiliki karakteristik yang berbeda atau sama dalam hal validitas, reliabilitas, tingkat kesukaran, maupun daya pembeda.

Mansyur, Soeratno, dan Harun (2015) mengemukakan bahwa masalah yang terjadi dalam praktik penilaian adalah dalam menafsirkan hasil pengukuran, dua atau beberapa perangkat tes sering diperlakukan sama, tanpa memperhatikan tingkat kesukaran perangkat tes yang digunakan. Masalah tersebut dapat diatasi dengan melakukan penyetaraan skor yang diperoleh dari peserta yang mengambil tes-tes itu. Sebagaimana dikemukakan oleh Miyatun & Mardapi (2000) tingkat kesetaraan perangkat tes yang berbeda akan dapat diketahui melalui proses penyetaraan.

Uraian di atas menggambarkan bahwa penyetaraan tes juga diperlukan oleh para guru di sekolah. Penyetaraan tes dirasakan kegunaannya mengingat mutu pendidikan (khususnya di Sulawesi Selatan) belum merata dengan keadaan geografis yang cukup luas. Tingkat kesetaraan tes dapat diperoleh melalui pengembangan tes yang setara. Namun demikian, tes tersebut dihadapkan pada perbedaan tingkat kesukaran dan perbedaan populasi yang dijadikan sasaran pelaksanaan. Oleh sebab itu, perlu dilakukan penyesuaian terhadap parameter-parameter tes dalam suatu skala yang sama dan penyesuaian skor tes dalam skala yang sama sehingga skor pada tes yang satu dapat dipertukarkan dengan skor pada tes lainnya.

Sukirno (2007) mengemukakan bahwa melalui proses penyetaraan diperoleh beberapa keuntungan, di antaranya *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. Jika kesetaraan paket tes sudah diketahui, pengukuran dapat dilakukan pada tempat dan waktu yang berbeda.

Dorans, Moses, dan Eignor (2010) mengatakan bahwa penyetaraan adalah bentuk kuat untuk menghubungkan antara skor pada dua tes. Tujuan penyetaraan adalah untuk menghasilkan skor pada dua bentuk tes sehingga skor dari setiap tes dapat diperbandingkan dari pengujian yang sama. Suatu keharusan bagi pengembang tes atau lembaga tes untuk menyetarakan perangkat tes tersebut.

Penyetaraan adalah proses statistik yang digunakan untuk mengatur skor pada format-format tes sehingga skor pada format tersebut dapat diperbandingkan (Kolen & Brennan, 2014). Hambleton, Swaminathan, dan Rogers (1991) menyatakan bahwa penyetaraan skor adalah membandingkan skor yang diperoleh dari perangkat tes yang satu

(X) dan perangkat tes lainnya (Y) yang dilakukan melalui proses penyetaraan skor pada kedua perangkat tes tersebut.

Proses penyetaraan dari beberapa perangkat tes (*equating*) dapat dilakukan dengan dua cara, yaitu penyetaraan secara horizontal dan penyetaraan secara vertikal (Croker & Algina, 2006). Proses penyetaraan yang diperoleh dari dua perangkat tes yang berbeda tetap mengukur hak yang sama dinamakan penyetaraan horizontal. Adapun proses penyetaraan dari dua kelompok peserta tes yang berbeda tingkat/jenjang pendidikannya, namun diberikan perangkat soal yang sama dinamakan penyetaraan vertikal.

Dalam pelaksanaannya, proses penyetaraan tes dilakukan berdasarkan pendekatan klasik dan modern. Untuk pendekatan klasik, proses penyetaraan tes digunakan teori "*true story*" dan untuk pendekatan modern digunakan teori respon butir (*Item Response Theory*) atau *Latent Trait Theory*. Proses penyetaraan dengan pendekatan klasik dapat dilaksanakan dengan mempergunakan tiga metode, yaitu (1) penyetaraan secara linier, (2) penyetaraan secara equipersentil, dan (3) penyetaraan secara curvalinear.

Lord (1980, p. 199) mengungkapkan tiga prinsip dasar untuk penyetaraan dua tes, sebagai berikut. (1) Kesetaraan (*equity*), untuk setiap kelompok peserta tes dengan kemampuan yang sama, kondisi distribusi frekuensi skor pada tes *Y* setelah transformasi adalah sama dengan distribusi frekuensi skor pada tes *X*. (2) *Population Invariance*, hubungan penyetaraan (transformasi) harus sama tanpa memperhatikan kelompok peserta tes (populasi) yang digunakan. (3) Simetri (*symmetry*), penyetaraan itu harus sama tanpa memperhatikan tes yang diberi label *X* atau diberi label *Y* atau transformasi dapat dibalik, artinya memetakan skor dari tes bentuk *X* ke tes bentuk *Y* sama dengan memetakan skor dari tes bentuk *Y* ke bentuk *X*.

Cook dan Eignor (1991) menyatakan bahwa prosedur penyetaraan dengan teori respon butir dapat dikelompokkan dalam tiga tahapan proses, yaitu: memilih desain pengumpulan data, penempatan parameter estimasi pada skala yang sama, dan penyetaraan skor tes. Selanjutnya, dalam desain *common-item nonequivalent group* yang juga dikenal dengan desain *anchor test* bahwa dua kelompok peserta tes yang berbeda, masing-masing memperoleh naskah tes yang berbeda pula, dan pada setiap naskah tes berisi kumpulan *anchor item* yang disebut dengan *anchor test*.

Anchor item adalah butir-butir soal yang sama di beberapa perangkat tes dan berbaur dengan butir yang *nonanchor*. Kelompok peserta tes tidak harus dipilih secara random dari populasi yang sama karena dalam praktiknya sering tidak sama. Hal tersebut sesuai dengan pendapat Kolen & Brennan (2014) bahwa desain *common-item nonequivalent group* menjelaskan kelompok peserta tes tidak harus dipilih secara random dari populasi yang sama dan di dalam praktik sering tidak sama. Kondisi tersebut merupakan salah satu keuntungan dari desain ini karena dalam keperluan praktik seringkali ditemui kondisi distribusi kemampuan kelompok berbeda.

Ketika desain *anchor item* digunakan, hendaknya memperhatikan sifat dan karakteristik dari *anchor item* dan penggunaan skornya. *Anchor item* harus menggambarkan miniatur tes yang disetarakan dan *item* tersebut relatif berada pada nomor urut yang sama, baik pada naskah tes yang pertama maupun naskah tes lainnya. Livingston sebagaimana dikutip oleh Hayati dan Mardapi (2014) menyatakan bahwa pertanyaan yang dimasukkan dalam butir *anchor* harus menggambarkan secara lengkap tingkat kesukaran dari butir soal,

dikarenakan hasil *equiting* tidak dapat tepat jika hanya memasukkan soal yang memiliki tingkat kesulitan tinggi atau soal yang memiliki tingkat kesulitan rendah. Sementara itu, jumlah *anchor item* disarankan 20% dari panjang tes untuk model tes yang diskor secara dikotomus (Kolen & Brennan, 2014).

Hasil penelitian berkenaan dengan jumlah *anchor item* yang digunakan pada penyetaraan tes model politomus telah dilakukan oleh Swediati (1997) yang menyimpulkan bahwa estimasi parameter yang rendah membuat sulit untuk menyamakan tes yang diberikan kepada kelompok ujian yang sangat berbeda dalam kemampuan, terutama jika tes yang relatif singkat dan jumlah item *anchor* kecil. Kolen dan Brennan (2014) menyatakan bahwa jumlah *anchor item* yang besar akan lebih mencerminkan tes dan ketika kedua tes yang sama diujikan kepada dua kelompok peserta tes, tidak perlu dilakukan penyetaraan karena skor tes yang dihasilkan berada dalam skala yang sama. Hal ini menggambarkan bahwa keseluruhan item dari kedua naskah tersebut berfungsi sebagai *anchor item*.Dengan demikian, ketika jumlah *anchor item* semakin mendekati jumlah butir dari suatu tes maka kualitas penyetaraan semakin akurat. Demikian juga dengan posisi *anchor item*, mempengaruhi kualitas penyetaraan sehingga posisi *anchor item* pada kedua naskah tes harus ditempatkan pada nomor item yang sama (Kolen & Brennan, 2014). Berdasarkan uraian di atas, pada penelitian ini dikaji tentang perbedaan akurasi hasil penyetaraan tes berdasarkan perbedaan jumlah *anchor item* (25%, 30%, dan 40%) dengan menggunakan teori respon butir.

METODE

Penelitian ini adalah penelitian eksploratif untuk mengungkapkan karakteristik soal fisika dengan menggunakan model logistik dua parameter (2PL) untuk kesetaraan tes dengan metode *equate-IRT* menggunakan *software* program R. Penelitian dilaksanakan selama dua bulan (Januari sampai dengan Februari 2016) pada enam Sekolah Menengah Atas (SMA) di Kabupaten Gowa Sulawesi Selatan, yaitu SMA Negeri 1 Sungguminasa, SMA Negeri 2 Sungguminasa, SMA Negeri 3 Sungguminasa, SMA Negeri 2 Tinggimoncong, SMA Negeri 1 Bajeng, dan SMA Negeri 1 Bajeng Barat.

Prosedur dalam penyetaraan tes terdiri dari beberapa tahap. Tahap *pertama*, pemilihan jenis penyetaraan. Dalam hal ini, jenis penyetaraan yang digunakan adalah penyetaraan horizontal, yakni penyetaraan yang dilakukan terhadap tingkat atau jenjang pendidikan yang sama (kelas XII IPA). Tahap *kedua*, pemilihan desain penyetaraan. Dalam hal ini, desain yang digunakan adalah desain *Common-Item Nonequivalent*. Desain *Common-Item Nonequivalent* ini merupakan desain yang menggunakan dua kelompok peserta tes yang berbeda dan dua perangkat tes yang berbeda, yaitu paket 01 dan paket 02 atau paket 03 dan paket 04 atau paket 05 dan paket 06. Kedua paket tersebut berisi kumpulan butir bersama atau yang disebut dengan *anchor item*. Jumlah *anchor item* yang digunakan adalah 10 butir atau 25% untuk soal paket 01 dan paket 02, 12 butir atau 30% untuk soal paket 03 dan paket 04, dan 16 butir atau 40% untuk soal paket 05 dan paket 06. Jumlah butir setiap paket adalah 40 butir tes. Tahap *ketiga*, dilakukan estimasi parameter model 2 parameter dengan menggunakan program *R* (*library ltm)*. Hal ini untuk melihat daya beda dan tingkat kesulitan pada paket 01 dan paket 02, paket 03 dan 04, dan paket 05 dan paket 06. Tahap *keempat*, proses *equating* atau penyetaraan.

Data dalam penelitian ini adalah respon (lembar jawaban) siswa SMA peserta tes di Kabupaten Gowa Tahun Pelajaran 2015/2016. Subjek penelitian terdiri atas 1420 orang (lembar jawaban) siswa dengan rincian sebagai berikut. *Pertama*, paket 01 sebanyak 240 orang peserta tes dan paket 02 sebanyak 235 orang peserta tes. *Kedua*, paket 03 sebanyak 246 orang peserta tes dan paket 04 sebanyak 239 orang peserta tes. *Ketiga*, paket 05 sebanyak 240 orang peserta tes dan paket 06 sebanyak 230 orang peserta tes. Data pasangan paket (misalnya paket 01 dan 02) pada program R dianalisis melalui *equateIRT*. Hasil analisis dari *equating* ini akan menunjukkan daya beda dan tingkat kesulitan dengan butir *anchor* pada masing-masing paket soal.

HASIL PENELITIAN DAN PEMBAHASAN

Analisis butir berdasarkan teori respon butir yang dilakukan dengan menggunakan *Program R versi. 3.2.2.* Analisis ini menggunakan model 2 parameter (2P) yang menghasilkan karakteristik butir yang meliputi tingkat kesulitan butir, daya pembeda butir, dan penyetaraan tes. Berdasarkan hasil pengolahan data yang telah dilakukan, karakteristik tingkat kesulitan dan daya pembeda masing-masing paket soal dielaborasi sebagai berikut.

Mencermati hasil pengolahan data ditinjau dari tingkat kesulitan soal, diperoleh untuk paket 01 sebanyak 2 butir (5%) tingkat kesukaran butir soal berada pada kategori sangat mudah, 3 butir (7,5%) kategori mudah, 23 butir (57,5%) pada kategori sedang, dan 12 butir (30%) kategori sukar. Untuk paket 02 diperoleh bahwa 2 butir (5%) tingkat kesukaran butir soal berada pada kategori sangat mudah, 3 butir (7,5%) kategori mudah, 30 butir (75%) kategori sedang, 4 butir (10%) kategori sukar, dan 1 butir (2,5%) kategori sangat sukar. Ditinjau dari daya pembeda butir, untuk paket 01 diperoleh daya pembeda yang bervariasi dengan rincian 38 butir (95%) tergolong sangat baik dan 2 butir (5%) berkategori baik. Pada paket 02 diperoleh daya pembeda dengan tiga variasi, yaitu 35 butir (87,5%) tergolong sangat baik, 1 butir (2,5%) tergolong baik, dan 4 butir (10%) tergolong buruk.

Berdasarkan hasil pengolahan data yang dilakukan diperoleh bahwa tingkat kesulitan soal untuk paket 03, yaitu 4 butir (10%) kategori sangat mudah, 3 butir (7,5%) kategori mudah, 26 butir (65%) kategori sedang, 6 butir (15%) kategori sukar dan 1 butir (2,5%) kategori sangat sukar. Untuk paket 04, tingkat kesulitan soal terdiri atas 3 butir (7,5%) kategori sangat mudah, 3 butir (7,5%) kategori mudah, 30 butir (75%) kategori sedang, 3 butir (7,5%) kategori sukar, dan 1 butir (2,5%) kategori sangat sukar dari 40 butir soal yang dianalisis. Karakteristik daya pembeda butir soal untuk paket 03 terdiri atas dua kategori, yaitu 38 butir (95%) kategori sangat baik dan 2 butir (5%) kategori buruk. Untuk paket 04, diperoleh daya pembeda yang bervariasi, yaitu 37 butir (92,5%) kategori sangat baik, 1 butir (2,5%) kategori baik, dan 2 butir (5%) kategori buruk.

Hasil analisis karakteristik paket soal 05 menunjukkan bahwa tingkat kesulitan butir soal diperoleh 3 butir (7,5%) kategori sangat mudah, 11 butir (27,5%) kategori mudah, 16 butir (40%) kategori sedang, 6 butir (15%) kategori sukar, dan 4 butir (10%) kategori sangat sukar. Untuk paket 06, terdapat 1 butir (2,5%) kategori sangat mudah, 3 butir (7,5%) kategori mudah, 27 butir (67,5%) kategori sedang, 5 butir (12,5%) kategori sukar, dan 1 butir (2,5%) kategori sangat sukar. Karakteristik daya pembeda butir soal untuk paket 05 dari 40 butir soal yang dianalisis, terdapat 37 butir (92,5%) kategori sangat baik dan 3 butir (7,5%) berkategori buruk. Untuk paket 06,

diperoleh daya pembeda yang bervariasi yaitu 35 butir (87,5%) kategori sangat baik, 2 butir (5%) kategori baik, dan 3 butir (7,5%) kategori buruk.

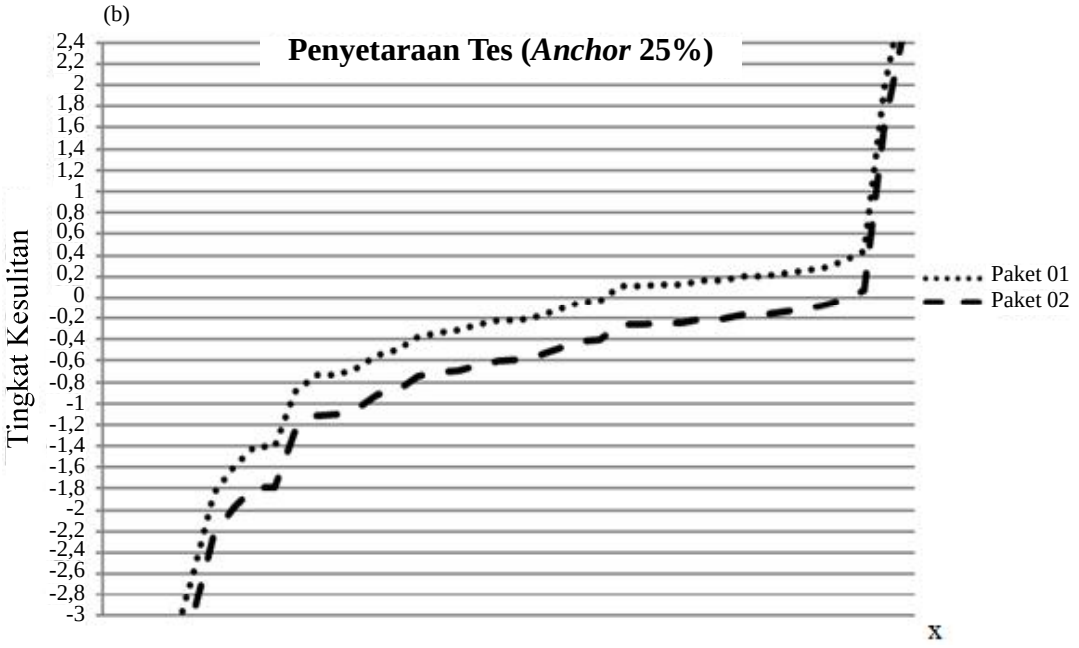
Sebagaimana dikemukakan sebelumnya, proses penyetaraan parameter butir soal paket 01dan paket 02 dengan banyaknya *anchor item* 25%, paket 03 dan 04 dengan *anchor item* sebanyak 30% serta paket 05 dan 06 dengan *anchor item*sebanyak 40%, dilakukan dengan menggunakan *library ltm (EquateIRT)* pada Program R dan model logistik 2 parameter. Parameter butir yang diperhatikan yakni tingkat kesulitan butir dan daya pembeda butir dengan butir *anchor* pada masing-masing paket soal.

Berdasarkan hasil penyetaraan paket 01 dan paket 02 dengan *anchor item* 25% dan metode *Haebara* pada *EquateIRT* terungkap bahwa kedua paket tersebut memiliki tingkat kesulitan yang berbeda. Paket 01 lebih sulit dari paket 02. Oleh karena itu, persamaan yang terbaik dengan tidak merugikan peserta didik adalah persamaan dari paket sulit ke paket mudah dengan persamaan b^*_{X2} = (1.00820). b_{X1}+(-0.36996). Dengan formula tersebut, diperoleh grafik penyetaraan tingkat kesulitan paket 01 dan paket 02 seperti yang disajikan pada Gambar 1.

Mencermati informasi yang disajikan pada Gambar 1, tampak bahwa hasil penyetaraan tingkat kesulitan antara paket 01 dengan paket 02. Garis paket 01 di atas garis paket 02. Hal ini menunjukkan bahwa paket soal dengan tingkat kesulitan rendah akan berada di bawah nilai kriteria karena proses *equating* yang dilakukan dari paket soal yang sulit ke paket soal yang mudah. Sebaliknya, paket soal dengan tingkat kesulitan tinggi akan berada di atas nilai yang menjadi kriteria. Dengan kata lain, proses *equating* dilakukan dari paket soal yang mudah ke paket soal sukar.

Ditinjau dari daya pembeda butir, hasil penyetaraan paket 01 dan paket 02

Gambar 1. Grafik Tingkat Kesulitan Paket 01 dan Paket 02 *Anchor Item* 25%



dengan *anchor* 25% metode Haebara pada *EquateIRT* terungkap bahwa soal paket 01 mampu membedakan kemampuan peserta yang tinggi dengan kemampuan peserta yang rendah. Soal pada paket 02 mampu membedakan kemampuan peserta yang tinggi dengan kemampuan peserta yang rendah, dengan persamaan (a*X2= a_X1/1.00820). Dengan formula tersebut, akan diperoleh grafik penyetaraan daya pembeda butir paket 01 dan paket 02 yang berhimpit. Artinya, hasil penyetaraan daya beda antara paket 01 dengan paket 02 berimpit. Hal ini menunjukkan bahwa kedua paket soal berada pada kategori tingkat daya beda yang sama. Pola garis kedua paket soal mengarah ke nilai positif, artinya kedua paket soal dapat membedakan peserta antara kemampuan yang tinggi dengan kemampuan rendah.

Berkaitan dengan penyetaraan soal paket 03 dan paket 04 dengan *anchor item* 30% terungkap bahwa kedua paket tersebut memiliki tingkat kesulitan yang berbeda. Paket 03 lebih sulit dibandingkan dengan paket 04. Oleh karena itu, persamaan yang terbaik dengan tidak merugikan peserta didik adalah persamaan dari paket sulit ke paket mudah dengan persamaan b*X4=(0.800241).bX3+(-0.077775). Dengan formula tersebut, diperoleh grafik penyetaraan tingkat kesulitan paket 03 dan paket 04 seperti disajikan Gambar 2.

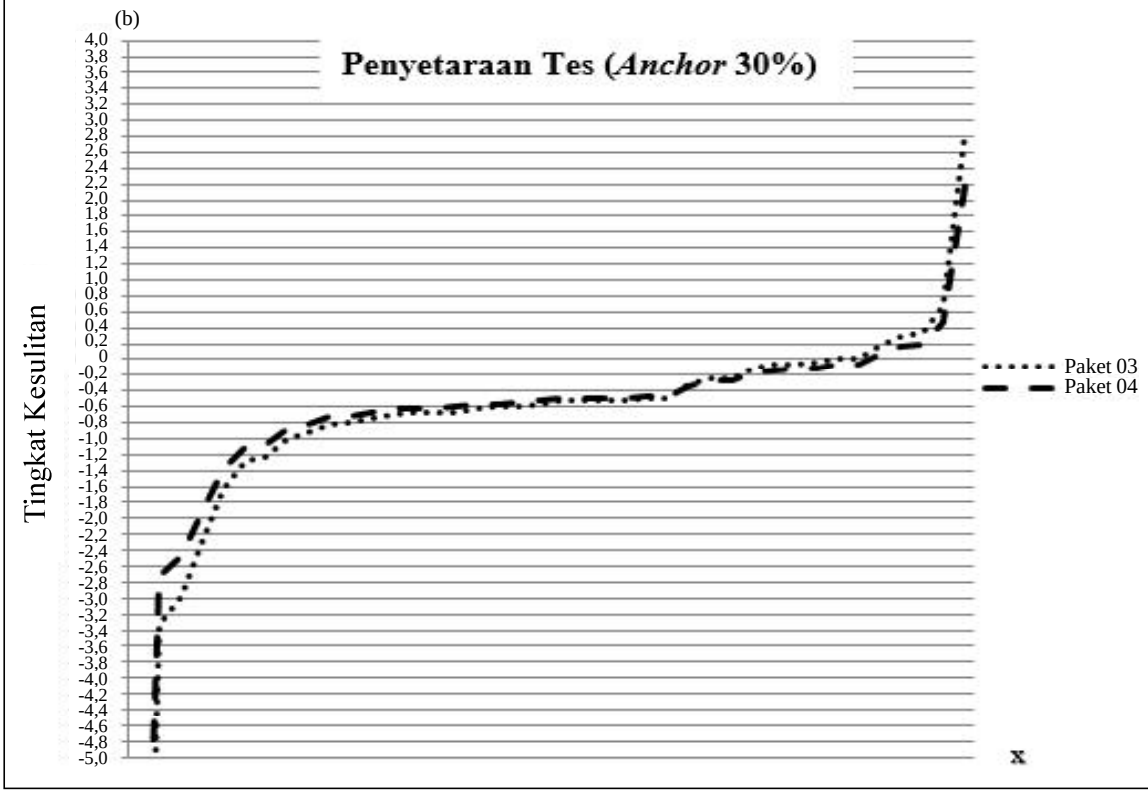
Mencermati informasi yang disajikan dalam Gambar 2, tampak hasil penyetaraan tingkat kesulitan antara paket 03 dengan paket 04. Garis paket 03 di atas garis paket 04. Hal ini menunjukkan bahwa paket soal dengan tingkat kesulitan rendah akan berada di bawah nilai kriteria karena proses *equating* yang dilakukan dari paket soal yang sulit ke paket soal yang mudah. Sebaliknya, paket soal dengan tingkat kesulitan tinggi akan berada di atas nilai yang menjadi kriteria. Artinya, proses *equating* yang dilakukakan dari paket soal yang mudah ke paket soal sukar.

Ditinjau dari daya beda, hasil penyetaraan tes paket 03 dan paket 04 dengan *anchor* item 30% terungkap bahwa kedua paket tersebut memiliki daya beda yang berbeda. Artinya, untuk dapat membedakan kemampuan tinggi dan rendah untuk kedua paket secara bersamaan sebaiknya menggunakan koefisien penyetaraan dengan formula (a*X4=a_X3/(0.800241)). Dengan formula tersebut, diperoleh grafik penyetaraan daya pembeda butir untuk paket 03 dan paket 04 tidak berhimpit. Posisi garis paket 03 lebih mengarah ke arah positif atau mendekati angka nol dibandingkan dengan paket 04. Hal ini menunjukkan bahwa paket 03 sangat baik dalam hal membedakan antara peserta tes kemampuan tinggi dengan kemampuan rendah dibandingkan dengan paket 04.

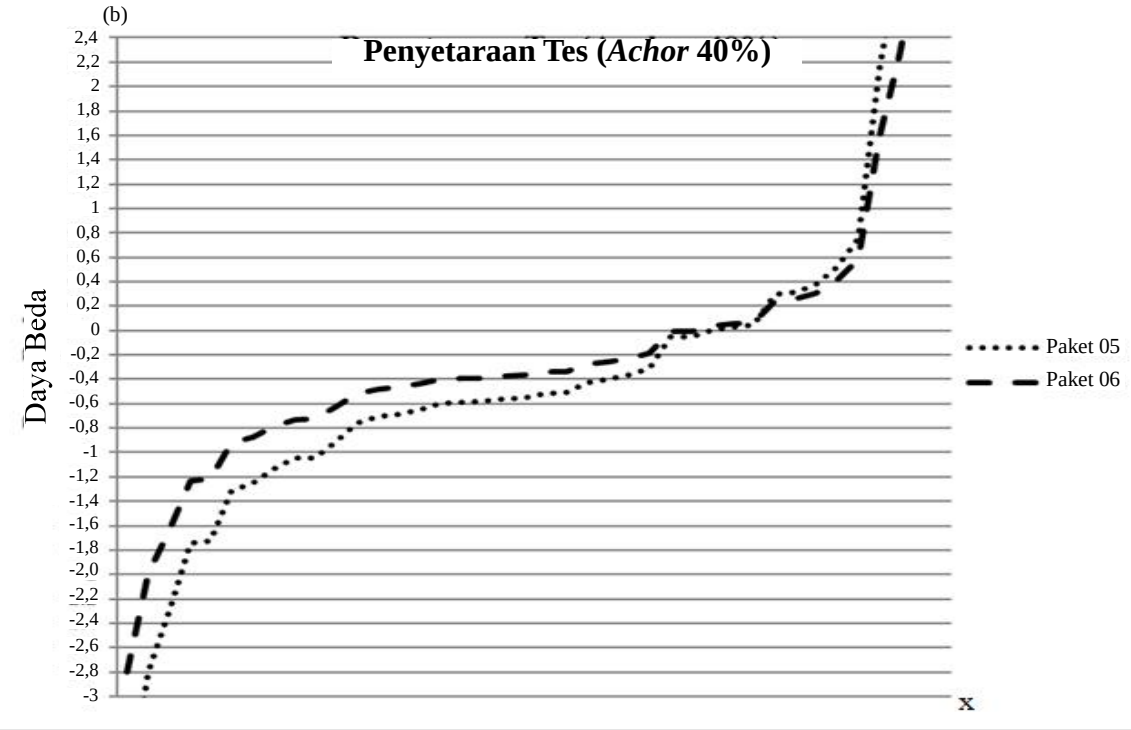
Hasil analisis penyetaraan soal paket 05 dan paket 06 dengan *anchor item* 40% dan metode Haebara pada *EquateIRT* terungkap bahwa kedua paket tersebut memiliki tingkat kesulitan yang berbeda. Paket 06 lebih sulit dibandingkan dengan paket 05. Oleh karena itu, persamaan yang terbaik dengan tidak merugikan peserta didik adalah persamaan dari paket sulit ke paket mudah dengan persamaanbX5= ((b_ X6-0.03269)/0.72467). Dengan formula tersebut, diperoleh grafik penyetaraan tingkat kesulitan paket 05 dan paket 06 seperti disajikan Gambar 3.

Informasi yang disajikan Gambar 3 menunjukan bahwa paket soal dengan tingkat kesulitan tinggi akan berada di atas nilai kriteria karena proses *equating* yang dilakukan dari paket soal yang mudah ke paket soal yang sulit. Sebaliknya, paket soal dengan tingkat kesulitan mudah akan berada di bawah nilai yang menjadi kriteria. Artinya, proses *equating* dilakukan dari paket soal yang sulit ke paket mudah.

Gambar 2. Grafik Tingkat Kesulitan Paket 03 dan Paket 04 *Anchor Item* 30%



Gambar 3. Grafik Tingkat Kesulitan Paket 05 dan Paket 06 *Anchor Item* 40%



Hasil penyetaraan paket 05 dan paket 06 dengan *anchor item* 40% dan metode Haebara pada *EquateIRT* terungkap bahwa daya beda pada paket 05 mampu membedakan kemampuan peserta yang tinggi dengan kemampuan peserta yang rendah. Paket 06 mampu membedakan kemampuan peserta yang tinggi dengan kemampuan peserta yang rendah. Kedua paket tersebut memiliki indeks daya beda yang berbeda. Oleh karena itu, dapat meng-gunakan koefisien penyetaraan dengan formula (a*X6=a_X5/(0.72467)). Dengan formula tersebut, diperoleh grafik penyetaraan daya pembeda butir untuk paket 05 dan paket 06 yang tidak berhimpit. Posisi garis paket 05 lebih mengarah ke arah positif atau mendekati angka nol dibandingkan dengan paket 06. Hal ini menunjukkan bahwa paket 05 sangat baik dalam hal membedakan antara peserta tes kemampuan tinggi dengan kemampuan rendah dibandingkan dengan paket 06.

Kualitas penyetaraan dilihat dari jumlah butir *anchor item* yang digunakan dalam penyetaraan menunjukkan bahwa semakin banyak jumlah *anchor item* yang digunakan pada proses penyetaraan mempengaruhi nilai koefisien penyetaraan. Secara teoretis, semakin banyak jumlah *anchor item* yang digunakan nilai koefisien α dan β semakin mendekati 1 dan 0. Demikian juga dengan nilai standar *error* dari koefisien penyetaraan, semakin banyak jumlah *anchor item* yang digunakan nilai standar *error* koefisien penyetaraan semakin kecil. Hal ini berarti semakin banyak jumlah *anchor item*, kualitas penyetaraan semakin akurat. Ringkasan hasil penyetaraan untuk masing-masing *anchor item* ditinjau dari koefisien α dan β disajikan pada Tabel 1.

Tabel 1 menunjukkan bahwa dari hasil analisis data terdapat perbedaan hasil penyetaraan antara paket soal dengan 10 butir *anchor* (25% dari 40 butir soal) dan paket soal dengan 12 butir *anchor* (30% dari 40 butir soal). Paket soal dengan 12 butir *anchor* menghasilkan koefisien penyetaraan yang lebih akurat dibandingkan paket soal dengan 10 butir *anchor* karena koefisien α pada paket soal dengan 12 butir *anchor* lebih mendekati 1 dan koefisien β lebih mendekati 0.

Tabel 1
Hasil Analisis Perbedaan Koefisien Penyetaraan Berdasarkan Perbedaan Jumlah Butir Anchor

Jumlah *Anchor*	Koefisien α	Koefisien β
10 *anchor* (25%)	1,00820	-0,36996
12 *anchor* (30%)	0,800241	-0,077775
16 *anchor* (40%)	0,72467	0,03269

Terdapat perbedaan koefisien penyetaraan pada paket soal dengan 10 butir *anchor* (25% dari 40 butir soal) dan paket soal dengan 16 butir *anchor* (40% dari 40 butir soal). Paket soal dengan 16 butir *anchor* menghasilkan koefisien penyetaraan yang lebih akurat dibandingkan paket soal dengan 10 butir *anchor* karena koefisien α pada paket soal dengan 16 butir *anchor* lebih mendekati 1 dan koefisien β lebih mendekati 0.

Terdapat perbedaan koefisien penyetaraan pada paket soal dengan 12 butir *anchor* (30% dari 40 butir soal) dan paket soal dengan 16 butir *anchor* (40% dari 40 butir soal). Paket soal dengan 16 butir *anchor* menghasilkan koefisien penyetaraan yang lebih akurat dibandingkan paket soal dengan 12 butir *anchor* karena koefisien α pada paket soal dengan 16 butir *anchor* lebih mendekati 1 dan koefisien β lebih mendekati 0. Ditinjau dari standar *error* yang dihasilkan dari penyetaraan untuk masing-masing paket soal disajikan pada Tabel 2.

Tabel 2
Hasil Analisis Standar Error Koefisien Penyetaraan Berdasarkan Perbedaan Jumlah Butir Anchor

Jumlah Anchor	Standar *Error*	
	Koefisien α	Koefisien β
10 *anchor* (25%)	0,21236	0,18060
12 *anchor* (30%)	0,16461	0,15494
16 *anchor* (40%)	0,11994	0,11748

Berdasarkan Tabel 2, terdapat perbedaan standar *error* hasil penyetaraan antara paket soal dengan 10 butir *anchor* (25% dari 40 butir soal) dan paket soal dengan 12 butir *anchor* (30% dari 40 butir soal). Paket soal dengan 12 butir *anchor* menghasilkan standar *error* yang lebih kecil dibandingkan paket soal dengan 10 butir *anchor*. Hal ini berarti hasil penyetaraan tes pada paket soal dengan 12 butir *anchor* lebih akurat dibandingkan dengan paket soal dengan 10 butir *anchor*.

Terdapat perbedaan standar *error* hasil penyetaraan antara paket soal dengan 10 butir *anchor* (25% dari 40 butir soal) dan paket soal dengan 16 butir *anchor* (40% dari 40 butir soal). Paket soal dengan 16 butir *anchor* menghasilkan standar *error* yang lebih kecil dibandingkan paket soal dengan 10 butir *anchor*. Hal ini berarti hasil penyetaraan tes pada paket soal dengan 16 butir *anchor* lebih akurat dibandingkan dengan paket soal dengan 10 butir *anchor*.

Terdapat pula perbedaan standar *error* hasil penyetaraan antara paket soal dengan 12 butir *anchor* (30 % dari 40 butir soal) dan paket soal dengan 16 butir *anchor* (40% dari 40 butir soal). Paket soal dengan 16 butir *anchor* menghasilkan standar *error* yang lebih kecil dibandingkan paket soal dengan 12 butir *anchor*. Hal ini berarti hasil penyetaraan tes pada paket soal dengan 16 butir *anchor* lebih akurat dibandingkan dengan paket soal dengan 12 butir *anchor*.

Hal ini sejalan dengan penelitian yang dilakukan sebelumnya pada data dikotomus (Hanson & Beguin, 2002) dan data politomus (Swediati, 1997). Hasil ini sesuai pula dengan yang telah diperkirakan atau dibahas pada kajian pustaka, seperti yang dikemukakan oleh Kolen & Brennan (2014) bahwa *anchor item* yang besar akan mencerminkan tes. Hasil ini juga sejalan dengan Battauz (2015, p. 101) yang menyatakan bahwa jumlah butir *anchor* memiliki pengaruh penting terhadap keragaman koefisien penyetaraan apabila ukuran sampel kecil terutama pada panjang tes. Liu, Sinharay, Holland, Curley, & Feigenbaum (2011) menyatakan bahwa hasil penyetaraan menunjukkan anchor kecil tidak selalu menghasilkan fungsi akurasi kesetaraan yang lebih baik dibandingkan *anchor* sedang. *Anchor* sedang yang dihasilkan menunjukkan sama baik atau bahkan lebih baik dari *anchor* kecil.

SIMPULAN

Berdasarkan hasil penelitian dan pembahasan di atas, dapat disimpulkan hal-hal sebagai berikut. *Pertama*, terdapat perbedaan hasil penyetaraan tes antara paket soal dengan 10 butir *anchor* (25% dari 40 butir soal) dan paket soal dengan 12 butir *anchor* (30% dari 40 butir soal). Paket soal dengan 12 butir *anchor* menghasilkan penyetaraan yang lebih akurat. *Kedua*, terdapat perbedaan hasil penyetaraan tes antara paket soal dengan 10 butir *anchor* (25% dari 40 butir soal) dan paket soal dengan 16 butir *anchor* (40% dari 40 butir soal). Paket soal dengan 16 butir *anchor* menghasilkan penyetaraan yang lebih akurat. *Ketiga*, terdapat perbedaan hasil penyetaraan tes antara paket soal dengan 12 butir *anchor* (35% dari 40 butir soal) dan paket soal dengan 16 butir *anchor* (40% dari 40 butir soal). Paket soal dengan 16

butir *anchor* menghasilkan penyetaraan yang lebih akurat. *Keempat*, paket soal dengan jumlah *anchor* yang paling besar menghasilkan penyetaraan tes yang lebih akurat. Berdasarkan simpulan tersebut, sebaiknya dilakukan penyetaraan tes berdasarkan jumlah butir *anchor* sebesar 40% agar kualitas penyetaraan tes yang diperoleh lebih akurat.

DAFTAR PUSTAKA

Battauz, M. (2015). Factors affecting the variability of IRT equating coefficients. *Statistica Neerlandica, 69*(2), 85-101.

Cook, L. L., & Eignor D. R. (1991). IRT equating methods. Educational testing service. *Educational Measurement: Issues and Practice, 10*, 37-45.

Croker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* California: Wadsworth Pub Co.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series, 2010*(2), i-41.

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

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common item non equivalent groups equating desain. *Applied Psychological Measurement, 26*, 3-34.

Hayati, N., & Mardapi, D. (2014). Pengembangan butir soal matematika SD di Kabupaten Lombok Timur sebagai upaya dalam pengadaan bank soal. *Jurnal Kependidikan, 44*(2), 26-38.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer Verlag Inc.

Liu, J., Sinharay, S., Holland, P. W., Curley, E., & Feigenbaum, M. (2011). Test score equating using a Mini-Version anchor and a midi anchor: A case study using SAT® data. *Journal of Educational Measurement, 48*(4), 361-379.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.

Mansyur, Soeratno, & Harun, R. (2015). *Asesmen pembelajaran di sekolah: Panduan bagi guru dan calon guru*. Yogyakarta: Pustaka Pelajar.

Miyatun, E., & Mardapi, D. (2000). Komparasi metode penyetaraan tes menurut teori respon butir. *Jurnal Penelitian dan Evaluasi, 2*(3), 1-18. Diunduh dari http://id.portalgaruda.org/?ref=browse&mod=viewarticle&article=168292.

Rosana, D., & Sukardiyono. (2015). Analisis butir dan identifikasi ketidakwajaran skor ujian akhir sekolah untuk standarisasi penilaian. *Jurnal Kependidikan, 44*(2), 92-102.

Sukirno, D. S. (2007). Penyetaraan tes UAN, mengapa dan bagaimana. *Jurnal Cakrawala Pendidikan, 26*(3), 305-321.

Swediati, N. (1997). *Equating tests under the generalized partial credit model* (Doctoral Dissertation). Diunduh dari http://scholarworks.umass.edu/dissertations/. (Order No. AAI9809405).

Syahrul. (2014). *Penerapan asesmen alteratif bagi peningkatan kualitas proses dan hasil belajar matematika siswa SMK Negeri 2 Makassar* (Laporan penelitian). Makassar: LPPM UNM.

Taruno, D. L. B. (2014). Model uji kompetensi keahlian instalasi listrik. *Jurnal Kependidikan,44*(2), 103-116.

INDEKS SUB-JEK