Jurnal Penelitian dan Evaluasi Pendidikan
Volume 19, No 2, Desember 2015 (176-188)

Tersedia Online: http://journal.uny.ac.id/index.php/jpep

PERBANDINGAN ESTIMASI KESALAHAN PENGUKURAN *STANDARD SETTING* DALAM PENILAIAN KOMPETENSI AKUNTANSI SMK

[1)]Sebastianus Widanarto Prijowuntato, [2)]Djemari Mardapi, [3)]Budiyono
[1)]Universitas Sanata Dharma, [2)]Universitas Negeri Yogyakarta, [3)]Universitas Negeri Surakarta
[1)]swidanartop@gmail.com, [2)]djemarimardapi@gmail.com, [3)]bud@uns.ac.id

Abstrak

Penelitian ini bertujuan untuk mengestimasi kesalahan pengukuran pada metode Angoff, Ebel, dan *Bookmark* dalam penilaian kompetensi Akuntnasi jenjang SMK di DIY yang digunakan *standard setter* dalam menentukan *cut score*. Penelitian ini merupakan penelitian kuantitatif. Sumber data dalam penelitian ini adalah respon peserta Ujian Nasional Praktik Akuntansi Paket 2 tahun ajaran 2011/2012 dengan 338 siswa. Guru-guru yang terlibat dalam *Focus Group Discussion* (FGD) berjumlah sembilan orang yang terdiri dari tujuh wanita dan dua pria. Teknik analisis dalam penelitian ini dibagi dalam tiga tahap yaitu: (1) persiapan, (2) FGD, (3) estimasi kesalahan pengukuran dengan menggunakan *Bootstrap*. Hasil penelitian menunjukkan bahwa *cut score* untuk metode Angoff sebesar 67,809, Ebel sebesar 59,034, dan *Bookmark* sebesar 57,022. Metode Angoff memiliki estimasi kesalahan pengukuran yang paling kecil (2,102) dibandingkan dengan metode Ebel (4,004) dan metode *Bookmark* (4,042). Oleh karena itu, metode Angoff merupakan metode yang tepat untuk mengestimasi kesalahan pengukuran pada *standard setting.*

Kata kunci: *Estimasi kesalahan pengukuran, Bootstrap, Cut Score*

ESTIMATION OF STANDARD SETTING ERROR MEASUREMENT IN ACCOUNTING COMPETENCY ASSESSMENT IN VOCATIONAL SCHOOLS

[1)]Sebastianus Widanarto Prijowuntato, [2)]Djemari Mardapi, [3)]Budiyono
[1)]Universitas Sanata Dharma, [2)]Universitas Negeri Yogyakarta, [3)]Universitas Negeri Surakarta
[1)]swidanartop@gmail.com, [2)]djemarimardapi@gmail.com, [3)]bud@uns.ac.id

Abstract

This research aims to estimate the measurement error in the Angof, Ebel, and Bookmark methods in Accounting Competency Assessment in Vocational Schools in DIY used by standard setters in deciding a cut score. This research is quantitative research. Data source in this study was the cut score of seven vocational schools in Yogyakarta that were randomly established. The reseach data were students' answers to the National Examination in Accounting Subject of Package 2 in the academic year of 2011/2012 with 338 students. The teachers who engaged in the Focus Group Discussion (FGD) were nine teachers, consisting of seven women and two men. The technical analysis was divided into three stages. 1) preparation, 2) FGD, 3) estimated error measurement by using the Bootstrap method. The results show that the cut score for the Angoff method is 67.809, Ebel method is 59.034, and Bookmark method is 57.022. The Angoff method has the least estimation of the measurement errors (2.102) as compared with the Ebel method (4.004) and the Bookmark method (4.042). Therefore, the Angoff method is the right method for estimating error measurement on standard setting.

Keywords: *Estimation of error measurement, Bootstrap, Cut Score*

Pendahuluan

Pembangunan di bidang pendidikan menjadi perhatian utama dari pemerintah. Pembangunan di bidang pendidikan ini tidak hanya mencakup aspek fisik seperti gedung-gedung, penyediaan sarana dan prasarana sekolah namun juga aspek non fisik seperti kurikulum, dan kualitas tenaga pendidik. Pembangunan ini bertujuan untuk meningkatkan kualitas pendidikan di Indonesia.

Terkait dengan kualitas pendidikan, pemerintah telah mengeluarkan standar pendidikan yang mencakup standar isi, standar kompetensi lulusan, standar sarana dan prasarana, standar pengelolaan, standar penilaian, standar proses, standar pembiayaan, standar pendidik dan tenaga kependidikan yang mencakup standar pengawas sekolah, standar kepala sekolah, standar kualifikasi akademik dan kompetensi guru. Standar-standar tersebut di atas merupakan standar yang harus dipatuhi oleh sekolah-sekolah. Pemerintah perlu menetapkan standar di bidang pendidikan karena standar dapat digunakan sebagai kriteria atau pembanding.

Kriteria atau pembanding digunakan untuk meningkatkan dan menyamakan kualitas pendidikan. Hal ini mengingat bahwa Indonesia memiliki wilayah yang cukup luas dan setiap daerah di wilayah Indonesia memiliki karakteristik yang berbeda-beda. Perbedaan karakteristik wilayah di Indonesia ini menjadi salah satu faktor penghambat dalam meningkatkan kualitas pendidikan. Dengan kata lain, masing-masing daerah di wilayah Indonesia memiliki kualitas pendidikan yang berbeda-beda. Pada umumnya, pendidikan di daerah perkotaan lebih maju daripada pendidikan di daerah pedalaman. Pemerintah perlu menyikapi perbedaan kualitas pendidikan tersebut secara bijaksana.

Di sisi lain, dalam berbagai bidang kehidupan, standar memiliki peran yang penting. Standar digunakan oleh masyarakat untuk menentukan baik atau buruk suatu produk. *Licensure, credentialing,* dan institusi akademik mencari pendekatan inovatif dalam *standard setting* untuk menilai kompetensi profesional (David, 2000, p.120). Di bidang kesehatan, masyarakat memerlukan angka kecukupan gizi untuk menentukan minimal gizi yang diperlukan agar dapat hidup sehat. Angka kecukupan gizi ini dicantumkan dalam berbagai kemasan makanan maupun minuman. Dalam bidang pekerjaan, institusi memerlukan standar untuk menentukan seseorang diterima atau tidak dalam pekerjaan tertentu. Dalam bidang psikologi, standar diperlukan untuk menggolongkan seseorang dalam kecerdasan (*intelligence quotient*) tertentu. Bidang pendidikan memerlukan standar untuk menentukan seseorang berhasil atau gagal dalam menempuh pendidikan tertentu. Standar pendidikan diperlukan juga untuk menyamakan kualitas pendidikan di seluruh wilayah Indonesia.

Perkembangan standar pendidikan ditandai dengan banyaknya penelitian-penelitian yang terkait dengan *standard setting* baik pembandingan metode (Koffler, 1980, p.6; Saunders, Ryan, & Huynh, 1980, p.2; Alsmadi, 2007, p.479; Skaggs, Hein, & Awuor, 2007, p.409; Retnawati, 2008, p.31; Premastuti, 2010, p.225), metode *standard setting* (Karantonis & Sireci, 2006, p.4), metode analisis (Chesser, et.al., 2004, p.825), estimasi *standard error* (Kane, & Wilson, 1984, p.107; Yin & Sconing, 2008, p.26; Le, 2000, p.605), validasi kinerja dengan menggunakan *passing score* (Kane, 1994, p.426). Perkembangan *standard setting* terus berlanjut seiring dengan perkembangan di bidang teknologi informasi.

Sampai saat ini, terdapat kurang lebih 50 metode yang digunakan untuk menentukan *standard setting* (Berk, 1986, p.137). Menurut Retnawati (2008, p.20), pada dasarnya metode-metode yang digunakan dalam *standard setting* dapat digolongkan menjadi empat golongan yaitu *standard setting* berdasarkan pada materi, butir/tes yang digunakan, peserta tes (*examinee*), dan kebijakan (*judgement*). Sementara itu, Livingstone & Zieky (1982, p.10) mengklasifikasikan *standard setting* ke dalam lima kelompok yaitu (1) metode yang berdasarkan pertanyaan tes (misalnya, *The Angoff method, The Angoff Mean Estimation Method, The Angoff Yes or No*

Method, The Nedelsky Method, The Bookmark Method, the Item Descriptor Matching Method), (2) metode yang berdasarkan pada profil skor (misalnya, *the Performance Profile Method, the Dominant Profile Method*), (3) metode yang berdasarkan pada pertimbangan orang-orang atau produk (misalnya, *the Borderline Group Method, the Contrasting Groups Method, the Up and Down Modification of the Contrasting Group Method, the Analytic Judgment Method*), (4) metode yang berdasarkan pada pertimbangan kelompok peserta tes (misalnya, *Judgment about a Reference Group, Judgment about Two Reference Groups*), (5) Metode yang berdasarkan pada kompromi antara pertimbangan absolut dan normatif (misalnya, *the Beuk Method, the Hofstee Method*). Penggolongan metode *standard setting* tersebut di atas dilakukan berdasarkan sudut pandang masing-masing ahli.

Cutscore merupakan salah satu isu yang penting dalam *standard setting*. Menurut Nudell (2008), penentuan *cut score* bukan hal yang mudah. Standar kompetensi minimal yang harus dicapai oleh seorang peserta tes harus ditentukan terlebih dahulu sebelum menentukan *cut score*. Apabila peserta dapat melampaui standar kompetensi minimal, maka peserta mencapai standar kompetensi yang dipersyaratkan untuk tujuan tertentu. *Cutscore* yang ditetapkan harus dapat mencerminkan ketercapaian kompetensi minimal yang harus dicapai oleh peserta tes.

Ketepatan dalam penentuan *standard setting* dapat dilihat dari besar kecilnya *error* dalam penentuan *standard setting*. Semakin besar *error* maka dapat dikatakan bahwa penentuan *cutscore* tidak tepat. Sebaliknya semakin kecil *error* maka penentuan *cutscore* semakin tepat. Livingstone & Zieky (2006, p.10) menyebutkan bahwa kemungkinan kesalahan menggunakan *cut score* terletak pada reliabilitas tes dan metode *cut score* yang digunakan. Dalam tes, tidak ada tes yang sangat reliabel. Tes yang sangat reliabel dapat ditunjukkan dengan nilai alpha Cronbach sebesar 1 (satu). Demikian juga dengan metode *cut score*, tidak ada metode *cut score* yang sempurna. Bila *cut score* yang ditentukan terlalu tinggi, maka siswa yang seharusnya lulus menjadi tidak lulus. Demikian pula sebaliknya, bila *cut score* ditentukan terlalu rendah, maka siswa yang seharusnya tidak lulus menjadi lulus. *Cut score* dapat dinaikkan maupun diturunkan. Konsekuensi dari menaikkan atau menurunkan *cut score* dapat menyebabkan *error* semakin membesar atau mengecil.

Metode *Bootstrap* digunakan untuk mengestimasi kesalahan pada pengukuran. Metode *Bootstrap* banyak diterapkan pada ilmu statistik untuk mengestimasi kesalahan pada populasi yang kecil atau populasi yang jumlahnya tidak diketahui. Asumsi yang mendasari metode *Bootstrap* adalah keindependenan datanya (Guan, 2003, p.72). Metode *Bootstrap* dilakukan dengan mengambil sampel dari populasi dengan pengembalian. Penelitian ini dilakukan untuk mengestimasi kesalahan pengukuran dengan menggunakan metode *Bootstrap*.

Di sisi lain, pendidikan SMK menyiapkan peserta didik untuk siap terjun di dunia usaha sesuai dengan standar pendidikan yang ada. Peningkatan kualitas pendidikan di SMK Program Studi Keahlian Keuangan Kompetensi Keahlian Akuntansi di samping mengikuti standar pendidikan yang sudah ditetapkan, juga mengikuti Standar Kompetensi Kerja Nasional Indonesia. Bagi pendidikan di SMK Keahlian Akuntansi, penetapan standar ini penting karena bidang pekerjaan lulusan SMK Keahlian Akuntansi diarahkan di bidang keuangan. Kesalahan yang terjadi dalam menginput satu transaksi saja akan mengakibatkan kesalahan pada laporan keuangan yang menjadi salah satu pertimbangan bagi perusahaan maupun *stakeholder* dalam pengambilan keputusan. Konsekuensinya, kurikulum pendidikan SMK Keahlian Keuangan ini disusun dengan memperhatikan standar pendidikan dan standar dunia usaha.

Diharapkan, lulusan SMK dapat langsung bekerja di dunia usaha sesuai dengan kompetensi yang dimiliki. Hal ini didukung oleh pola pendidikan di SMK yang lebih banyak menekankan pada praktik kerja industri di samping mendapatkan materi di kelas. Praktik kerja industri bertujuan agar

siswa memiliki bekal dalam mempraktikan materi-materi yang diperoleh di kelas. Permasalahan yang dihadapi dalam praktik kerja industri adalah tidak banyak industri mau memberikan pelatihan kepada siswa sesuai dengan bidang keahliannya.

Permasalahan yang dihadapi dalam dunia pendidikan di Indonesia adalah penetapan *cutscore* sebagai penentu kelulusan belum didasarkan atas salah satu metode *standard setting* dan belum pernah diestimasi besarnya kesalahan penentuan *cut score*.

Kesalahan pengukuran muncul ketika kerangka konseptual yang menganggap bahwa konstruk yang diukur adalah invarian atas beberapa kondisi pengamatan (Kane, 2010, p.7). Dengan kata lain dapat dikatakan bahwa terdapat deviasi dari sesuatu diukur dengan konstruk yang menyusunnya. Teori pengukuran berkaitan dengan pengembangan tolok ukur atau instrumen dengan bantuan seorang analis sistem atau peneliti yang dapat mengukur atribut suatu entitas/fenomena/sistem yang diteliti (Chadha, 2009, p.5). Dalam pengukuran, tidak semua atribut yang diukur/diwakili secara sempurna mewakili konstruknya.

Setiap pengukuran yang dilakukan akan memiliki kesalahan (*error*). Mardapi (2008, p.2) mengungkap bahwa kesalahan yang terjadi pada pengukuran di bidang ilmu alam lebih sederhana dibandingkan dengan ilmu sosial. Kesalahan pengukuran pada ilmu alam lebih disebabkan karena alat ukurnya, sedangkan kesalahan pengukuran pada ilmu sosial dapat diakibatkan karena alat ukurnya, cara pengukurannya, dan objek yang diukur.

Senada dengan Mardapi (2008, p.2), Nichols, Twing, & Mueller (2010, p.15) menyatakan bahwa masalah yang dihadapi proses pengukuran dalam ilmu sosial adalah tidak adanya indikator yang langsung dapat mengukur atribut yang akan diukur. Oleh karena itu, ilmu sosial menggunakan struktur data untuk mengukur atribut laten. Untuk memecahkan masalah ini, maka Nichols, Twing, & Mueller, (2010, p.15) mengajukan tiga pendekatan yaitu (1) *subject-centered approach*, (2) *stimulus-centered approach*, dan (3) *response-centered approach*. *Subject-centered approach* bertujuan untuk menskalakan orang sehubungan dengan attribut laten. *Stimulus-centered approach* bertujuan untuk menskalakan rangsangan. Sedangkan resp*onse-centered approach* bertujuan untuk menskalakan orang, stimuli atau keduanya.

Kesalahan pengukuran dapat diterima apabila kesalahan pengukuran tersebut merupakan kesalahan yang paling minimal. Kesalahan yang paling kecil menunjukkan bahwa objek yang diukur memiliki kesesuaian dengan konstruknya dan dapat dikatakan bahwa alat ukur tersebut handal. Alat ukur yang handal akan memberikan hasil yang konsisten apabila alat tersebut digunakan berulang-ulang.

Cohen, Kane, & Crooks (1999, p.359) menyebutkan bahwa kesalahan dalam *cut score* bersumber pada dua hal yaitu *error inherent in the estimation of the equating relations dan the sampling error associated with the selection of rater*. *Error inherent in the estimation of the equating relations* terjadi ketika proses penetapan standar dilakukan secara berulang-ulang oleh penilai yang sama. Setiap kali perulangan penilaian, penilai beristirahat sejenak dan mendiskusikan hasil penilaiannya. Hal ini menyebabkan hasil penilaian berikutnya tidak independen karena penilai telah terpengaruh oleh hasil diskusi yang dilakukan. *The sampling error associated with the selection of rater* terjadi ketika rater berbeda menilai pekerjaan siswa. Hasil penilaian *rater* yang satu akan berbeda dengan penilaian hasil rater yang lainnya. Selanjutnya, Cohen, Kane, & Crooks (1999, p.360) merumuskan *standard error measurement* (SEM) sebagai dengan rumus sebagai berikut.

$$SEM(C) = \frac{\delta_c}{\sqrt{n}}$$

C adalah *cut score* ke C, δ_c merupakan deviasi standar estimasi dan n merupakan banyaknya replikasi *cut score*.

Estimasi *error* pengukuran dapat dilakukan dengan mendasarkan pada *generalized analysis of covariance structure model* yang dikembangkan oleh Jöreskog (1974, dalam Whitely, 1979, p.144). Pada dasarnya, model Jöreskog berusaha untuk mereproduksi

matriks kovarians dengan rumus struktural $\sum = \wedge\Phi\wedge' + \psi^2$. Dalam model kovarians umum, urutan dari $\wedge, \Phi,$ dan ψ lebih kecil atau sama dengan urutan dari Σ. Tiga tipe dari parameter mungkin terkandung dalam $\wedge, \Phi$ dan $\psi^2 -$ parameter bebas, tetap, atau terbatas. Parameter tetap dan terbatas adalah spesifikasi model.

Feldt, Steffen, & Gupta (1985, p.353) mengajukan lima macam metode untuk mengestimasi kesalahan baku yaitu sebagai berikut.

Pertama teori klasik yang dikemukakan oleh Lord & Novic. Estimasi kesalahan baku dengan menggunakan teori klasik diperoleh dengan menggunakan rumus sebagai berikut:

$S_E = S_x(1 - r_{xx'})^{1/2}$.

Kedua, Pendekatan Thorndike. Pendekatan Thorndike mendasarkan pada teori tes klasik. Estimasi kesalahan baku dengan menggunakan pendekatan Thorndike diperoleh dengan menggunakan rumus sebagi berikut.

$X_1 = T_1 + E_1$.

Ketiga, Pendekatan Polinomial. Dengan menggunakan pendekatan ini, kesalahan estimasi diperoleh dengan menggunakan rumus sebagai berikut:

$\hat{Y} = a_0 + a_1(X) + a_2(X^2) + a_3(X^3) + a_4(X^4)$.

Keempat, Pendekatan binomial Lord: *modification* Keats. Pada pendekatan ini, individual *i* dipandang mampu menjawab proporsi tertentu, Φ, dari keseluruhan butir. Dengan konseptualiasi ini, skor *examinee i* pada satu bentuk tes analog dengan perhitungan frekuensi dari kejadian fenomena Q dalam *random sample* dari unit-unit *k*. Rumus $S_E = \left[\left(\frac{X(k-X)}{k-1}\right)\left(\frac{1-r_{xx^1}}{1-r_{21}}\right)\right]^{1/2}$ digunakan untuk mengestimasi kesalahan pengukuran.

Kelima, Pendekatan binomial Lord: *compound binomial*. Formula yang digunakan dalam pendekatan ini adalah sebagai berikut.

$S_{E(i)} = \left[\sum_{h=1}^{C} \frac{X_{ih}(k_h - X_{ih})}{k_h - 1}\right]^{1/2}$.

Cowel, W.R. (1991, p.2) menguraikan empat metode untuk mengestimasi kesalahan baku pada pengukuran. Keempat metode tersebut adalah (1) *Mached half test (split halves)*, (2) *Item response theory* (IRT), (3) *Randomly parallel forms (binomial)*, (4) *Matched parallel form (compound binomial)*.

Estimasi *error* pengukuran juga dapat dilakukan dengan mendasarkan pada *generalized analysis of covariance structure model* yang dikembangkan oleh Jöreskog (1974, dalam Whitely, 1979, p.143). Pada dasarnya, model Jöreskog berusaha untuk mereproduksi matriks kovarian dengan rumus struktural: $\sum = \wedge \Phi \wedge' + \psi^2$

Estimasi *error* pengukuran juga dapat didekati dengan metoda *Bootstrap*. *Bootstrap* merupakan pendekatan nonparametrik yang mengijinkan seseorang untuk menghindar dari perhitungan teoritis (Guan, 2003, p.72). Hal ini berdasaran asumsi bahwa sampel yang ada merupakan representasi dari populasi. Dengan menggunakan *Bootstrap,* estimator bias, varians, dan statistik lainnya dapat dihitung. Cara yang dilakukan untuk mengestimasi populasi dengan menggunakan *Bootstrap* adalah dengan mengambil sampel dari data asli dan mengembalikannya lagi.

Kesalahan baku pada metode *Bootstrap* dapat dicari dengan menggunakan rumus sebagai berikut (Efron, & Tibshirani, 1993. p.47).

$$\widehat{se}_B = \left\{\sum_{b=1}^{B}\left[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)\right]^2/(B-1)\right\}^{1/2}$$

Dengan $\hat{\theta}^* =$ parameter populasi, $B =$ *banyaknya resampel yang dilakukan,* dan $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B}\hat{\theta}^*(b)/B$ merupakan nilai mean sederhana dari $\hat{\theta}^*(1)$, $\hat{\theta}^*(2) \ldots \hat{\theta}^*(B)$

Dalam penelitian ini, metode *Bootstrap* digunakan untuk mereplikasi *cut score* yang diperoleh melalui metode Angoff, Ebel, dan Bookmark. *Cut score* merupakan fokus dalam penetapan standar. Kinerja seseorang akan selalu dikaitkan dengan standar minimal yang ditetapkan. Seperti kelulusan ujian, penerimaan pegawai, penerimaan kredit, dan sebagainya. Dapat dikatakan bahwa *cut score* merupakan titik batas yang dapat menggolongkan peserta tes ke dalam beberapa golongan seperti lulus atau tidak lulus, atau *basic, proficient,* dan *advance*. Penggolong-

an *examinee* ini tergantung pada kepentingan dan tujuan diselenggarakannya tes. Dengan demikian setiap bidang, setiap negara memiliki *cut score* yang tidak sama.

Bejar (2008, p.2) mendefinisikan *cut score* sebagai skor yang mengklasifikasikan siswa yang belum mencapai *score* level 1 dan siswa yang memiliki skor di atas 1. *Cut score* yang digunakan untuk memisahkan kemampuan siswa merupakan hasil dari *Standard setting*. Dalam konteks *standard setting*, *cut score* merupakan salah satu komponen kunci.

Metode Penelitian

Penelitian ini merupakan penelitian kuantitatif. Sumber data yang digunakan dalam penelitian ini adalah respon peserta didik SMK dalam Ujian Nasional Praktik Akuntansi tahun pelajaran 2011/2012 sebanyak 338 siswa di Daerah Istimewa Yogyakarta. Guru-guru yang dilibatkan dalam FGD berjumlah sembilan orang yang memenuhi persyaratan sebagai berikut: (1) ahli dalam bidang yang berhubungan dengan ujian; (2) terbiasa dengan metode-metode ujian; (3) dapat memecahkan masalah dengan baik; (4) terbiasa dengan level kandidat; dan (5) tertarik dengan pendidikan (guru); (6) telah mengajar minimal 5 tahun; (7) mengajar di kelas 12 dan 8) lulusan dari program studi Ekonomi/Akuntansi. Teknik analisis dalam penelitian ini dibagi ke dalam tiga tahap. Tahap pertama yaitu persiapan. Kegiatan pada tahap ini mencakup penyiapan data, penggolongan SMK, dan meneliti karakteristik butir dengan menggunakan program Quest dengan 1 parameter logistik (1PL). Tahap kedua yaitu FGD. Sebelum dilakukan FGD, guru-guru dilatih dalam menentukan *cut score*. Hal ini bertujuan agar guru memiliki persepsi yang sama dalam menentukan *cut score*. FGD dilakukan dalam dua putaran. Dalam putaran pertama, peserta FGD diberi pelatihan dalam menentukan *cut score* dengan menggunakan metode Angoff, Ebel, dan *Bookmark*. Pada putaran kedua, peserta FGD menentukan *cut score* namun tidak diberi pelatihan lagi. Tahap ketiga yaitu mengestimasi kesalahan pengukuran dengan menggunakan *Bootstrap*. Estimasi kesalahan pengukuran dilakukan dengan menggunakan program R.

Hasil Penelitian dan Pembahasan

Tingkat kesulitan butir pada penelitian ini dianalisis dengan menggunakan program Quest. Tingkat kesulitan butir untuk 74 butir (6 butir jurnal, 39 butir buku besar, dan 9 siklus akuntansi) pada penelitian ini diperoleh dengan menggunakan program *Quest*. Tingkat kesulitan pada output *Quest* dapat dilihat pada *DIFF*. Hasil analisis *Quest* untuk butir jurnal disajikan pada Tabel 1.

Tabel 1. Tingkat Kesulitan Butir Jurnal

Kriteria	No Butir	Jumlah	%
Mudah (<-2)	2, 10, 23	3	11,54
Baik (-2 ≤x≤2)	1, 3, 5, 6, 8, 9, 13, 14, 15, 16, 17, 18, 19, 21, 22, 24, 26	17	65,38
Sulit (>2)	4, 7, 11, 12, 20, 25	6	23,08
Jumlah		26	100

Tingkat kesulitan butir pada buku besar diketahui terletak di antara -4 sampai 4. Jumlah soal yang tergolong mudah pada butir Buku Besar adalah 3 butir (7,69%), butir yang tergolong baik berjumlah 32 butir (82,05%), dan butir yang termasuk dalam golongan butir yang sulit berjumlah 4 butir (10,26%). Keseluruhan tingkat kesulitan butir disajikan pada Tabel 2.

Tabel 2. Tingkat Kesulitan Butir Buku Besar

Kriteria	No Butir	Jumlah	%
Mudah (<-2)	17, 23, 33	3	7,69
Sedang (-2 ≤x≤2)	01, 02, 04, 06, 07, 08, 09, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 25, 26, 27, 28, 29, 30, 31, 32, 34, 35, 36, 38, 39	32	82,05
Sulit (>2)	03, 05, 24, 37	4	10,26
Jumlah		39	100

Pada butir-butir siklus akuntansi tidak terdapat soal yang mudah. Butir yang tergolong baik berjumlah 7 butir (77,78%), dan soal yang tergolong sulit berjumlah 2 butir (22,22%). Tabel 3 menunjukkan rangkuman butir menurut tingkat kesulitan butir.

Tabel 3. Tingkat Kesulitan Butir Siklus Akuntansi

Kriteria	No Butir	Jumlah	%
Mudah (<-2)	-	-	0
Sedang (-2 ≤x≤2)	01, 02, 03, 04, 05, 06, 07	7	77,78%
Sulit (>2)	08, 09	2	22,22%
Jumlah		9	100%

Pelaksanaan penentuan *Cutscore* diawali dengan memberikan paparan terkait dengan pengertian dan tujuan *standard setting*, instrumen yang digunakan, dan analisis data dengan menggunakan metode Angoff, Ebel, dan Bookmark. Bahan-bahan diskusi yang dibagikan kepada peserta terdiri dari penjelasan tentang *standard setting*, *round down standard setting*, lembar jawaban panelis untuk masing-masing metode *standard setting*, standar kompetensi lulusan (SKL), soal ujian praktik kejuruan, rubrik penilaian, dan *ordered item booklet* (OIB).

Setelah penjelasan tentang materi *standard setting*, peserta berlatih *standard setting* di bawah bimbingan peneliti. Setelah peserta berlatih, nilai-nilai yang ada disimulasikan dengan menggunakan program Excell. Setelah berlatih *standard setting*, peserta kemudian melakukan *standard setting* putaran pertama dan putaran kedua untuk masing-masing metode.

Cut score yang diperoleh dengan menggunakan metode Angoff, Ebel, dan *Bookmark* tampak pada Tabel 4. Data pada tabel 4 kemudian dijadikan sebagai populasi untuk *Bootstrap*. Sampel yang diambil secara dari populasi tersebut di atas berjumlah tujuh buah. Hasil pengambilan sampel tampak pada Tabel 5.

Tabel 4. Data *Cut Score*

Panelis	Putaran 1			Putaran 2		
	Angoff	Ebel	*Bookmark*	Angoff	Ebel	*Bookmark*
1	72,00	58,48	51,30	72,00	58,53	51,00
2	68,00	33,89	69,04	67,00	37,00	69,17
3	72,00	68,05	73,33	73,00	69,25	76,58
4	73,00	68,20	76,50	72,00	68,20	73,71
5	70,00	69,80	78,50	70,00	69,00	78,50
6	64,00	59,70	47,29	60,00	59,20	50,88
7	54,00	50,80	58,00	60,00	53,10	58,00
8	73,00	57,45	59,96	73,00	59,15	54,98
9	70,00	70,41	57,04	67,00	68,39	41,88
Rerata	68,44	59,64	63,44	68,22	60,20	61,63
Std dev	6,13	11,78	11,28	5,19	10,51	13,17

Tabel 5. Sampel *Cut Score*

Panelis	Putaran 1			Putaran 2		
	Angoff	Ebel	*Bookmark*	Angoff	Ebel	*Bookmark*
1	72,00	58,48	51,30	72,00	37,00	51,00
2	72,00	68,05	69,04	73,00	68,20	69,17
3	70,00	68,20	73,33	72,00	69,00	73,71
4	64,00	59,70	78,50	60,00	59,20	50,88
5	54,00	50,80	47,29	60,00	53,10	58,00
6	73,00	57,45	58,00	73,00	59,15	54,98
7	70,00	70,41	57,04	67,00	68,39	41,88

Pengambilan sampel dengan pengembalian (x^*) dilakukan pada masing-masing sampel *cut score* (x) untuk masing-masing metode. Pengambilan sampel tersebut dilakukan sebanyak 200 kali dan pada masing-masing pengambilan dicari standard deviasinya. *Bootstrap* tersebut dilakukan dengan menggunakan program R i386 3.0.0. Tabel 6 menunjukkan kesalahan pengukuran yang terjadi ketika pengambilan sampel dilakukan sebanyak 200 kali.

Tabel 6. Kesalahan Pengukuran

Keterangan		Rata-rata *cut score* (200 kali)	Kesalahan Pengukuran *Bootstrap* (200 kali)
Putaran 1	Angoff	67,903	2,429
	Ebel	61,960	2,419
	Bookmark	61,960	4,267
Putaran 2	Angoff	67,809	2,102
	Ebel	59,034	4,004
	Bookmark	57,022	4,042

Cut score yang tinggi ini dapat menggambarkan kemampuan/kompetensi siswa yang sebenarnya dalam mengelola akuntansi. Hal ini mengingat bahwa dalam dunia usaha, akuntansi memegang peranan yang cukup penting. Kesalahan yang terjadi pada tahap-tahap akuntansi akan mengakibatkan perusahaan merugi bahkan gulung tikar. Demikian pula sebaliknya, ketepatan dalam mengelola akuntansi menjadikan perusahaan dapat menyusun strategi dengan baik, menggaji karyawan tepat waktu, melakukan diversifikasi usaha, dan sebagainya.

Pencapaian kompetensi akuntansi yang tinggi tersebut menuntut perbaikan pada kurikulum, sarana dan prasarana, proses pendidikan, kualitas guru. Kurikulum yang digunakan dalam SMK perlu mengacu pada kebutuhan dunia usaha. Konsep *link and match* dalam pendidikan kejuruan menyelaraskan kebutuhan dunia usaha dengan dunia pendidikan. Agar konsep *link and match* dapat berjalan dengan baik, maka perlu ada komunikasi antara dunia usaha dan dunia pendidikan.

Dalam praktik, terkadang terjadi ketidakselarasan antara dunia usaha dan dunia pendidikan. Sebagai contoh, ketika siswa melakukan praktik kerja lapangan (PKL) banyak dunia usaha yang mengalihkan siswa program keahlian akuntansi ke bidang lainnya seperti sekretariat, tenaga pemasaran dan sebagainya. Banyak perusahaan yang menganggap bahwa keuangan merupakan rahasia perusahaan dan tidak boleh sembarang orang mengetahui.

Sarana dan prasarana pendidikan yang dimiliki oleh sekolah perlu ditingkatkan. Berdasarkan survei yang dilakukan oleh peneliti, banyak sekolah yang tidak memiliki sarana dan prasarana yang memadai untuk pembelajaran akuntansi. Banyak sekolah yang memiliki kelas terbatas yang mengakibatkan sekolah tidak memiliki laboratorium akuntansi. Di samping itu alat-alat pembelajaran yang ada di sekolah kejuruan sangat minim.

Minimnya sarana dan prasarana yang ada akan berakibat pada proses pembelajaran yang kurang optimal dan berdampak pada rendahnya kualitas lulusan kejuruan. Oleh karena itu, di samping guru dituntut lebih kreatif dalam membelajarkan akuntansi kepada siswa, guru juga perlu memperkenalkan praktik-praktik akuntansi dan membawanya di dalam kelas. Oleh karena itu, kreativitas guru perlu ditingkatkan dengan memberikan pendidikan dan pelatihan yang baik.

Kondisi-kondisi tersebut di atas perlu menjadi perhatian dan pertimbangan pemerintah meningkatkan kualitas pendidikan. Pemerintah hendaknya perlu memberikan perhatian lebih kepada sekolah-sekolah yang kualitasnya masih di bawah standar. Kesenjangan antara dunia usaha dan dunia pendidikan dapat diperkecil dengan adanya koordinasi antara kementerian yang terkait. Pemerintah juga perlu memberikan kontrol terhadap sekolah-sekolah khususnya sekolah kejuruan baik administrasi, tenaga pendidik dan kependidikan, proses pembelajaran melalui dinas pendidikan di kabupaten/kota.

Besarnya estimasi kesalahan pengukuran untuk setiap metode *standard setting* berbeda-beda. Pada putaran pertama untuk 200 kali *Bootstrap*, estimasi kesalahan pengukuran pada metode Angoff sebesar 2,429 dengan mean sebesar 67,903, metode Ebel sebesar 2,419 dengan mean sebesar 61,960, dan metode *Bookmark* sebesar 4,267 dengan mean sebesar 61,960. Pada putaran kedua untuk 200 kali *Bootstrap*, estimasi kesalahan pengukuran pada metode Angoff sebesar 2,102 dengan mean sebesar 67,809, metode Ebel sebesar 4,004 dengan mean sebesar 59,034, dan metode *Bookmark* sebesar 4,042 dengan mean sebesar 57,021.

Hasil penelitian ini berbeda dengan penelitian yang dilakukan oleh Premastuti, N.B. (2010, p.230) yang menyatakan bahwa metode *Bookmark* lebih akurat dibandingkan dengan metode *Group Contrast*. Menurut Premastuti (2010, p.230), metode *Bookmark* lebih baik karena dalam prosedurnya mempertimbangkan (1) parameter tingkat kesukaran butir dan (2) estimasi panelis dalam mengestimasi tiap butir dengan respon probabilitas tertentu.

Hasil penelitian ini juga berbeda dengan penelitian yang dilakukan oleh Widayati (2009, p.182). Widayati (2009, p.182) menyatakan bahwa metode teori respon butir merupakan metode yang paling tepat untuk mengestimasi kesalahan pengukuran perangkat soal uji coba ujian nasional mata pelajaran Biologi SMA tahun pelajaran 2007/2008.

Ada beberapa penyebab perbedaan hasil penelitian ini dengan dua penelitian di atas. Pertama, penyebab yang berkaitan dengan sampel. Sampel dalam penelitian yang dilakukan oleh Premastuti (2010, 225) dan Widayati (2009, p.192) merupakan sampel yang diambil dari populasi. Ketika sampel yang diambil besar, maka hasilnya dapat menggambarkan parameter populasi. Dalam penelitian ini, sampel yang digunakan bukan diambil dari populasi melainkan *resample* (Efron & Tibshirani, 1993, p.45). Kesalahan pengukuran pada penelitian ini akan menunjukkan hasil yang berbeda apabila jumlah sampel dan resampelnya ditambah. Mengingat jumlah sampel yang digunakan dalam penelitian ini tergolong sedikit, maka variasi resampelnya juga relatif sedikit. Jika sampel penelitian ditambah, maka *resample*-nya akan sangat bervariasi dan menghasilkan kesalahan pengukuran yang berbeda.

Perbedaan besaran kesalahan pengukuran pada metode *Bootstrap* disebabkan karena jumlah sampel yang digunakan relatif sedikit. Jumlah sampel akan sangat mempengaruhi variasi data dalam penelitian. Apabila sampel yang ada dilakukan pengambilan ulang dengan pengembalian, maka data yang diperoleh tidak memiliki banyak variasi.

Hasil estimasi kesalahan pengukuran diperoleh dengan cara mengambil sampel (x^*) dari sampel asli (x). *Resample* tersebut dilakukan sebanyak 200 kali dengan bantuan program R. Hasil *resample* untuk setiap kali *Bootstrap* akan berbeda, namun hasil kesalahan pengukuran tidak jauh berbeda.

Kedua, penyebab yang berkaitan dengan penilai. Besar/kecilnya estimasi kesalahan pengukuran tergantung pada kemampuan penilai, dalam hal ini adalah guru, dalam memprediksi kemampuan siswa. Guru yang memahami kemampuan siswa akan dengan mudah memprediksi kemampuan siswa dengan tepat. Pemahaman guru terhadap kemampuan siswa terjadi ketika guru sering berdiskusi dengan siswa tentang mata pelajaran ataupun pengetahuan yang lainnya.

Hasil perhitungan estimasi kesalahan pengukuran pada metode Angoff lebih kecil daripada metode Ebel dan *Bookmark*. Hal ini dikarenakan penentuan *cut score* pada metode Angoff lebih mudah daripada dua metode lainnya. Pada metode Angoff, guru diminta untuk mengestimasi kemampuan siswa dalam menjawab pertanyaan yang diujikan. Pada metode Ebel, guru diminta untuk memprediksi tingkat kesulitan, tingkat relevansi butir dan menentukan proporsi siswa yang mampu menjawab butir pada tiap sel. Sedangkan pada metode *Bookmark*, guru diminta untuk mengestimasi kemampuan siswa terlebih dahulu sebelum menentukan batas kelulusan.

Metode penelitian yang dilakukan Widayati (2009, p.192) tidak melibatkan guru dalam mengestimasi kesalahan pengukuran. Widayati, W. menggunakan data berupa respon siswa terhadap perangkat soal Biologi tahun pelajaran 2007/2008. Data tersebut digunakan untuk mengestimasi kesalahan baku pengukuran dengan menggunakan enam metode.

Dari kedua penyebab tersebut diduga menyebabkan hasil yang berbeda baik dalam pemilihan metode *standard setting* maupun dalam pemilihan metode kesalahan pengukuran. Untuk memberikan dukungan bahwa metode *Bootstrap* memberikan hasil estimasi kesalahan pengukuran yang lebih kecil, maka dilakukan pembandingan hasil estimasi kesalahan pengukuran dengan menggunakan metode *Keats*' *Modification*.

Hasil estimasi kesalahan pengukuran pada beberapa metode *standard setting* dengan menggunakan metode *Keats' Modification* tampak pada Tabel 7. Putaran pertama, metode *Keats' Modification* menghasilkan estimasi kesalahan pengukuran metode Angoff sebesar 3,292 dengan mean sebesar 67,903, metode Ebel sebesar 4,418 dengan

mean sebesar 61,960, dan metode *Bookmark* sebesar 4,418 dengan mean sebesar 61,960. Pada putaran kedua, estimasi kesalahan pengukuran pada metode Angoff sebesar 3,314 dengan mean sebesar 67,809, metode Ebel sebesar 4,809 dengan mean sebesar 59,034, dan metode *Bookmark* sebesar 5,033 dengan mean sebesar 57,022.

Tabel 7. Perbandingan Kesalahan Pengukuran *Bootstrap* dengan Keats' Modification

Keterangan		Kesalahan Pengukuran *Bootstrap* (200 kali)	Kesalahan Pengukuran Keats' Modification
Putaran 1	Angoff	2,429	3,292
	Ebel	2,419	4,418
	Bookmark	4,267	4,418
Putaran 2	Angoff	2,102	3,314
	Ebel	4,004	4,809
	Bookmark	4,042	5,033

Pada putaran 2, metode *Keats' Modification* menunjukkan bahwa rerata estimasi kesalahan pengukuran pada metode Angoff lebih kecil (3,314) daripada metode Ebel (4,809) dan metode *Bookmark* (5,033). Oleh karena itu, metode Angoff menunjukkan metode yang lebih baik untuk menentukan *cut score* karena memberikan estimasi kesalahan pengukuran yang lebih kecil dibandingkan dengan metode Ebel dan *Bookmark*.

Pada Tabel 7 tampak bahwa metode *Bootstrap* dan *Keats' Modification* menunjukkan hasil yang sama. Kedua metode tersebut menunjukkan bahwa metode Angoff merupakan metode standard setting yang lebih tepat dibandingkan metode Ebel dan *Bookmark* karena menghasilkan *standard error* yang paling kecil. Dengan demikian, kedua metode dapat digunakan untuk mengestimasi kesalahan pengukuran.

Kedua metode estimasi kesalahan standar tersebut menggunakan asumsi yang berbeda. Metode *Bootstrap* mendasarkan asumsi pada *central limit theorm* (CTL) sedangkan metode *Keats' Modification* mendasarkan pada distribusi binomial. CTL berkaitan dengan distribusi sampel yang diambil. Semakin banyak sampel yang diambil dari populasi maka distribusi sampelnya akan mendekati normal. Pada metode *Bootstrap*, semakin banyak *Bootstrap* dilakukan, maka hasil *Bootstrap* akan mendekati populasinya.

Distribusi binomial merupakan distribusi dari n percobaan berhasil/gagal yang saling bebas. Pada distribusi binomial, urutan observasi merupakan kejadian independen, dan probabilitas sukses dinyatakan dengan "p" (Subiyakto, 1995, p.45). Distribusi binomial seringkali digunakan untuk memodelkan jumlah keberhasilan pada jumlah sampel n dari populasi N.

Perhitungan estimasi kesalahan pengukuran *Keats' Modification* mempertimbangkan probabilitas jawaban benar dan probabilitas jawaban salah. Di samping itu, rumus *Keats' Modification* juga mempertimbangkan reliabilitas soal. Hal ini berbeda dengan SEM. Pada SEM, proporsi jawaban siswa dan reliabilitas soal tidak dipertimbangkan. Namun demikian, apabila dilihat dari kesalahan pengukuran yang dihasilkan, metode *Bootstrap* menghasilkan kesalahan pengukuran yang lebih kecil dibandingkan dengan metode *Keats' Modification*.

Berdasarkan hasil estimasi kesalahan pengukuran, metode Angoff merupakan metode yang lebih baik untuk menentukan *cut score* dibandingkan metode Ebel maupun *Bookmark*. Hal tersebut ditunjukkan oleh hasil estimasi kesalahan pengukuran pada Angoff lebih kecil (2,102 dengan menggunakan *Bootstrap*, dan 3,314 dengan menggunakan *Keats' Modification*) dibandingkan dengan metode Ebel (4,004 dengan menggunakan *Bootstrap*, dan 4,809 dengan menggunakan *Keats' Modification*) maupun *Bookmark* (4,042 dengan menggunakan *Bootstrap*, dan 5,033 dengan menggunakan *Keats' Modification*). Hasil ini sejalan dengan pendapat dari Anto & Mardapi, (2013, p.377) yang menyatakan bahwa bias estimasi pada Yes/No Angoff lebih kecil dibandingkan dengan metode Ebel.

Di samping itu, hasil estimasi kesalahan pengukuran, ketepatan penggunaan

metode *standard setting* juga ditunjukkan pada kemudahan panelis dalam menentukan *cut score*. Metode Angoff dan *Bookmark* merupakan meode yang mudah untuk diterapkan dibandingkan dengan metode Ebel. Pada metode Ebel, panelis melakukan estimasi terkait dengan tingkat kesulitan butir, tingkat relevansi, dan tingkat kemampuan peserta tes. Banyaknya estimasi yang dilakukan oleh panelis dapat menyebabkan tingkat kesalahan yang dilakukan semakin banyak. Berbeda dengan metode Angoff dan *Bookmark*, pada kedua metode tersebut panelis hanya mengestimasi tingkat kemampuan peserta tes. Namun demikian, pada metode *Bookmark*, ada beberapa indikator yang diestimasi dapat dikerjakan oleh siswa tidak tercakup dalam halaman *Bookmark* (Rejeki, Mardapi, & Kumaidi, 2014, p.94). Sementara pada metode Angoff, seluruh indikator kemampuan diestimasi oleh panelis.

Simpulan dan Saran

Simpulan

Berdasarkan hasil penelitian dan pembahasan yang telah diungkapkan maka dapat disampaikan beberapa simpulan sebagai berikut.

Pertama, *cut score* mata pelajaran Akuntansi jenjang SMK di DIY yang dihasilkan dengan menggunakan metode Angoff pada putaran 1 adalah 67,86 mengalami kenaikan yang tidak signifikan pada putaran 2 menjadi 68,14. Kedua, *cut score* mata pelajaran Akuntansi jenjang SMK di DIY yang dihasilkan dengan menggunakan metode Ebel pada putaran 1 adalah 61,87 mengalami penurunan yang tidak signifikan putaran 2 menjadi 59,15. Ketiga, *cut score* mata pelajaran Akuntansi jenjang SMK di DIY yang dihasilkan dengan menggunakan metode *Bookmark* pada putaran 1 adalah 62,07 mengalami penurunan signifikan pada putaran 2 menjadi 57,03.

Keempat, kesalahan pengukuran *cut score* mata pelajaran Akuntansi Jenjang SMK di DIY dengan menggunakan *Bootstrap* pada metode Angoff pada putaran 1 untuk Uji Kompetensi Kejuruan Praktik Akuntansi sebesar 2,429. Kesalahan pengukuran pada putaran pertama mengalami penurunan yang tidak signifikan sebesar 0,327 menjadi 2,102 pada putaran kedua. Sementara itu, kesalahan pengukuran dengan menggunakan metode Keats' Modification pada metode Angoff putaran pertama sebesar 3,153 mengalami kenaikan yang tidak signifikan sebesar 0,059 menjadi 3,212 pada putaran kedua.

Kelima, kesalahan pengukuran *cut score* mata pelajaran Akuntansi Jenjang SMK di DIY dengan menggunakan *Bootstrap* pada metode Ebel pada putaran 1 untuk Uji Kompetensi Kejuruan Praktik Akuntansi sebesar 2,419. Kesalahan pengukuran pada putaran pertama mengalami kenaikan yang tidak signifikan sebesar 1,586 menjadi 4,004 pada putaran kedua. Sementara itu, kesalahan pengukuran dengan menggunakan metode Keats' Modification pada metode Ebel putaran pertama sebesar 4,734 mengalami penurunan yang tidak signifikan sebesar 0,072 menjadi 4,663 pada putaran kedua.

Keenam, kesalahan pengukuran *cut score* mata pelajaran Akuntansi Jenjang SMK di DIY dengan menggunakan *Bootstrap* pada metode *Bookmark* pada putaran 1 untuk Uji Kompetensi Kejuruan Praktik Akuntansi sebesar 4,267. Kesalahan pengukuran pada putaran pertama ini mengalami penurunan yang tidak signifikan sebesar 0,225 menjadi 4,042 pada putaran kedua. Sementara itu, kesalahan pengukuran dengan menggunakan metode Keats' Modification pada metode *Bookmark* putaran pertama sebesar 4,187 mengalami kenaikan yang tidak signifikan sebesar 0,280 menjadi 4,467 pada putaran kedua.

Ketujuh, estimasi kesalahan pengukuran *cut score* yang paling tepat adalah metode Angoff karena memberikan rerata estimasi kesalahan pengukuran yang paling kecil.

Saran

Berdasarkan hasil penelitian dan beberapa keterbatasan penelitian maka saran yang dapat disampaikan sebagai berikut.

Pertama, penentuan *cut score* pada penelitian ini hanya terkait dengan Ujian Keterampilan Kejuruan Praktik Akuntansi. Untuk memberikan gambaran yang sesungguhnya, maka penilaian sikap, proses, kerapian, kecepatan dan kebersihan perlu dimasukkan dalam *standard setting*. Kedua, perlu dilakukan penelitian lebih lanjut terkait dengan *cut score* tentang penguasaan materi akuntansi di sekolah menengah kejuruan (aspek pengetahuan).

Ketiga, perlu dilakukan penelitian lain dengan menggunakan metode *standard setting* yang berbeda dan menambah panelis yang berasal dari dunia usaha/dunia industri sehingga menghasilkan penelitian yang lebih baik. Keempat, jumlah panelis perlu ditambah lebih dari sembilan orang agar dapat menggambarkan *cut score* yang lebih baik mengingat panelis dalam penelitian ini berjumlah 9 orang. Kelima, guru perlu memberikan latihan-latihan soal praktik yang bervariasi dan terprogram agar siswa benar-benar dapat menguasai praktik Akuntansi dan meningkatkan kualitas pembelajaran dengan menggunakan berbagai metode pembelajaran yang bervariasi.

Daftar Pustaka

Alsmadi, A. A. (2007). A comparative study of two standard-setting technique. *Social Behavior and Personality*, 38 (4), 479 – 486.

Anto, S., & Mardapi, D. (2013). *Komparasi metode standard setting untuk penentuan KKM mata pelajaran Matematika kelas VIII SMP*. Jurnal Penelitian dan Evaluasi Pendidikan. 17 (2). 369 – 388.

Bejar, I. I. (2008). Standard setting: what is it? why is it important?. *R&D Connection*, 7, 1 – 5.

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research,* 56, (1), 137 – 172.

Chadha, N.K. (2009). *Applied psychometric*. First Publishing, India: Vivek Mehra for Sage Publication.

Chesser, A. M. S., Laing, M. R., Miedzybrodzka, Z., Brittenden, J., et al (2004). Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment. *Medical Education*, 38, 825 – 831.

Cohen, A.S., Kane, M.T., & Crooks, T.J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, 12(4), 343 – 366.

Cowel, W.R. (1991). A procedure for estimating the conditional standard error of measurement for GRE general and subject test. *GRE Board Professional Report* No. 87-03P, ETS Research Report, 91 – 25.

David, B. (2000). AMEE Guide No. 18: Standard setting in student assessment, *Medical Teacher*, 22 (2), 120 – 130.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall. Inc.

Feldt, L.S., & Steffen, M., & Gupta C. N. (1985). A comparison of five methods for estimating the standard error of measurement at specific score model. *Applied Psycological Measurement*, 9 (4), 351 – 361.

Guan, W. (2003). From the help desk: bootstrapped standard errors. *The Stata Journal*. 3 (1), 71 – 80.

Kane, M.T., & Wilson, J. (1984). Errors of measurement and standard setting in mastery testing. *Applied Psychological Measurement*. 8 9 (1), 107 – 115.

Kane, M.T. (1994). Validating the performance standards associated with passing score. *Review of Educational Reseach*. 64 (3), 425 – 461.

Kane, M.T. (2010). *Error of measurement, theory, and public policy*. Educational Testing Service.

Karantonis, A,. & Sireci, S. G. (2006), The Bookmark standard-setting method: a literatur review, *Educational Measurement: Issues and Practice*, Spring, 4 - 12

Koffler, S. L. (1980). A Comparison of appoaches for setting proficiency standards, *Journal of Education Measurement*, 17 (3), 167 – 178.

Lee, G. (2000, 24 – 28 April). *Estimating reliability and standard error of measurement for complex reading comprehension tests under generalizability theory model*. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA.

Livingstone, S. A., & Zieky, M. J. (1982). *Passing scores: a manual for setting standards of performance on educational and occupational tests*, Princeton, New Jersey: Educational Testing Service.

Livingstone, S. A., & Zieky, M. J. (2006). *A manual for setting standards of performance on educational an occupational tests*. Princeton, New Jersey: Educational Testing Service.

Mardapi, D. (2008). *Teknik Penyusunan instrumen tes dan non tes*. Cetakan Pertama.Yogyakarta: Mitra Cendikia Press,

Nichols, P., Twing, J., & Mueller, C.D. (2010). Standard-setting methods as measurement process. *Educational Measurement: Issues and Practice,* 29 (1), 14 – 24.

Nudell, H. (2008, February). Making the cut - the *cut score*, that is establishing a pass/fail score is a highly technical process. *ICSC Certified Professionals Newsletter*.

Premastuti, N. B. (2010). Komparasi standard setting metoda group contrast dan Bookmark pada mata pelajaran Akuntansi. *Jurnal Penelitian dan Evaluasi Pendidikan*, 14 (2), 225 – 245.

Rejeki, S., Mardapi, D., & Kumaidi. (2014). Metode standard setting untuk ujian nasional di sekolah dasar. *Jurnal Penelitian dan Evaluasi Pendidikan. 18 (1), 89 – 97*.

Retnawati, H. (2008). *Penentuan batas lulus (standard setting) ujian nasional mata pelajaran Matematika di DIY*. Laporan Penelitian. UNY: PKPSP LP.

Saunders, J.C., Ryan, J.P., & Huynh, H. (1980, March 5 – 9). *A comparison of two ways of setting passing scores based on the nedelsky procedure*. Publication Series in Mastery Testing. South Carolina: University of South Carolina College of Education Colombia. this Article is presented at the annual conference of the Eastern Educational Research Association, Norfolk, Virginia.

Skaggs, G., Hein, S.F., & Awuor, R. (2007). Setting passing Scores on passage-based test: a comparison of traditional and single-passage Bookmark method, *Applied Measurement In Education*. 20 (4), 405 – 426.

Subiyakto, H. (1995). *Statistika (inferen) untuk bisnis*. Edisi ke-1. Cetakan ke-1. Yogyakarta: Bagian Penerbitan sekolah Tinggi Ilmu Ekonomi.

Whitely, S. E. (1979). Estimating measurement error on highly speeded tets. *Applied Psychological Measurement*, 3 (2), 141-154.

Widayati, W. (2009). Komparasi beberapa metode estimasi kesalahan pengukuran. *Jurnal Penelitian dan Evaluasi Pendidikan*. 13 (2). 182 – 197.

Yin, P., & Sconing, J., (2008). Estimating standard errors of cut scores for item rating and mapmark procedure: a generalizability theory approach. *Educational and Psychological Measurement*, 68 (1). 25 – 41.