Jurnal Penelitian dan Evaluasi Pendidikan
Volume 22, No 1, June 2018 (24-34)
Online: http://journal.uny.ac.id/index.php/jpep

ESTIMASI KESALAHAN PENGUKURAN *STANDARD SETTING* DALAM PENILAIAN KOMPETENSI MATEMATIKA TINGKAT SMP DI KABUPATEN SUMBAWA

Weni Wendari [1] *, *Samsul Hadi* [1]
[1]Prodi Penelitian & Evaluasi Pendidikan Program Pascasarjana Universitas Negeri Yogyakarta
[1]Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
* Corresponding Author. Email: wendari.weni@gmail.com

Abstrak

Penelitian ini bertujuan untuk mengetahui metode yang lebih akurat dalam mengestimasi kesalahan pengukuran *standard setting* pada metode *Ebel*, *Bookmark*, dan *Contrasting group*. Data penelitian ini merupakan dokumen Dinas Pendidikan dan Kebudayaan Kabupaten Sumbawa berupa respon peserta Ujian Nasional Matematika Paket P0C5520 tahun ajaran 2015/2016 yang berjumlah 352 siswa. Gsuru juga dilibatkan dalam penelitian sebagai panelis dalam *Focus Group Discussion* (FGD). Data yang terkumpul kemudian dianalisis melaui tiga tahap.. Tahap pertama yaitu persiapan, kegiatan pada tahap ini mencakup penyiapan data, penggolongan SMP, dan penentuan karakteristik butir. Tahap kedua yaitu FGD dilakukan dalam dua putaran. Tahap ketiga yaitu mengestimasi kesalahan pengukuran dengan menggunakan pendekatan *Generalizability Theory* dengan bantuan program *eduG*. Hasil penelitian menunjukkan bahwa metode *Contrasting group* memiliki estimasi kesalahan pengukuran paling kecil dibandingkan metode *Ebel* dan *Bookmark*, oleh karena itu, metode *Contrasting group* lebih akurat dibandingkan dengan dua metode lainnya.

Kata kunci: *cut score, standard setting, generalizability theory*

MEASUREMENT ERROR ESTIMATION OF STANDARD SETTING IN MATHEMATICS COMPETENCY ASSESSMENT FOR JUNIOR HIGH SCHOOL IN SUMBAWA REGENCY

Abstract

This research aims to find the most accurate methods in estimating measurement error of standard setting among Ebel, Bookmark, and Contrasting group methods. The data used in this study were 352 students' responses on Mathematics National Exam Package P0C5520 in the academic year of 2015/2016. The document was collected from the Department of Education and Culture in Sumbawa Regency. Teachers were also involved in this research as panelists in the Focus Group Discussion (FGD). The data collected were then analyzed through three stages. The first stage was preparation stage, including the activities of data preparation, school classification, and item characteristics analysis. The second stage was two-round FGD. The third stage was estimating the measurement error using Generalizability Theory approach assisted by eduG program. The research result shows that Contrasting Group method produces the smallest measurement error estimation compared to Ebel and Bookmark methods, therefore, Contrasting group method is considered as the most accurate method.

Keywords: *cut score, standard setting, generalizability theory*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v22i1.16492

Pendahuluan

Penilaian dalam kurikulum didesain dengan menggunakan kriteria tertentu. Asumsi yang digunakan pada kriteria yaitu bahwa setiap peserta didik dapat belajar pelajaran apa saja, namun dengan membutuhkan waktu yang berbeda. Kriteria itu berlaku untuk semua peserta didik tanpa membedakan mata pelajaran. Hasil penilaian sering dipandang sebagai tolak ukur keberhasilan peserta didik dalam proses pembelajaran. Hasil penilaian berdasarkan kriteria dikategorikan menjadi dua yaitu lulus dan tidak lulus. Peserta didik dikategorikan lulus jika memenuhi kriteria yang telah ditentukan. Sebaliknya, peserta didik dinyatakan tidak lulus apabila tidak memenuhi kriteria yang telah ditentukan.

Penentuan kriteria kelulusan dapat dilakukan melalui *judgment*. Namun penetapan dengan cara ini memiliki kelemahan, yakni tidak didasarkan pada data empirik dan prosedur yang telah teruji di lapangan. Penetapan dengan *judgment* ini hanya dilakukan berdasarkan pertimbangan dan pendapat yang tidak didasarkan pada data empirik. Kemudian pendapat dan pertimbangan tersebut dijadikan sebagai kebijakan. Sehingga kriteria kelulusan yang ditetapkan dengan cara ini tidak dapat digunakan karena tidak merepresentasikan keadaan yang sebenarnya di lapangan.

Saat ini, Indonesia menggunakan Kurikulum 2013 yang mana kriteria kelulusan untuk Ujian Nasional tidak ditentukan oleh pemerintah. Permendikbud Nomor 5 Tahun 2015 tentang Kriteria Kelulusan Peserta Didik dalam Ujian Nasional (UN) (Menteri Pendidikan dan Kebudayaan Republik Indonesia, 2015) menyatakan bahwa kelulusan peserta didik ditetapkan oleh satuan pendidikan, dalam hal ini kriteria kelulusan ditentukan oleh masing-masing sekolah atau yang biasa dikenal dengan otonomi sekolah. Berarti bahwa setiap sekolah di Indonesia mempunyai kriteria kelulusan yang berbeda–beda, misalnya di Kabupaten Sumbawa terdapat 101 sekolah menengah pertama (SMP) yang terdiri atas 95 SMP Negeri dan 6 SMP Swasta. Masing–masing SMP di Kabupaten Sumbawa memiliki kriteria kelulusan yang berbeda. Dimana kriteria kelulusan ditentukan hanya berdasarkan *intake* siswa, daya dukung, dan kompleksitas. Ketiga subskor tersebut selanjutnya diambil rata-rata yang kemudian digunakan sebagai batas kelulusan. Penggunaan teknik ini tentu menimbulkan masalah yang cukup serius terutama berkenaan dengan masalah reliabilitas atau keakuratan. Penentuan ketiga komponen tersebut memberikan konsekuensi akan tingginya variabilitas nilai yang mungkin muncul dari para penilai. Permasalahan reliabilitas ini dapat dieliminasi jika batas kelulusan atau kriteria kelulusan ditentukan dengan menggunakan *standard setting*.

Standard setting merupakan proses penentuan suatu titik atau batas dalam skala skor tes tertentu yang digunakan untuk menentukan level performa suatu kebijakan atau untuk membuat klasifikasi (Cizek, 1996, p. 20). *Standard setting* dalam dunia pendidikan banyak digunakan untuk menentukan skor batas kelulusan atau biasa disebut sebagai skor minimum kelulusan. Namun lebih dari itu, *standard setting* juga dapat digunakan sebagai alat bantu untuk memetakan mutu pendidikan, sebab dengan *standard setting* dapat dibuat suatu klasifikasi kompetensi seseorang atau prestasi suatu wilayah tertentu yang termasuk kategori tinggi, sedang, dan rendah. Tidak hanya digunakan dalam menentukan skor minimum kelulusan, *standard setting* juga dapat dimanfaatkan dalam memetakan mutu pendidikan dengan mengklasifikasikan kompetensi peserta didik atau prestasi suatu wilayah tertentu.

Secara garis besar, *standard setting* dibedakan menjadi dua golongan. Pertama yang menggunakan acuan norma dan kedua menggunakan acuan kriteria. Interpretasi nilai beracuan norma yaitu kemampuan peserta tes dibandingkan dengan kemampuan orang lain dalam kelompok acuan. Sementara interpretasi nilai beracuan kriteria, kemampuan peserta tes dibandingkan dengan level kemampuan tertentu.

Sampai saat ini, terdapat sekitar 38 metode yang digunakan dalam menentukan *standard setting* (Berk, 1986, p. 151). Metode

standard setting dibagi kedalam lima kelompok (Zieky, Perie, & Livingston, 2008, p. 86) yaitu (1) metode berdasarkan pertanyaan tes terdiri dari metode *Angoff*, metode estimasi rata–rata, metode *Yes or No Angoff*, metode *Nedelsky*, metode *Ebel*, metode *Bookmark*, dan metode *Item Descriptor Matching*; (2) metode berdasarkan profil skor terdiri dari metode profil performan, dan metode profil *dominant*; (3) metode berdasarkan pada pertimbangan orang atau produk terdiri dari metode *borderline group*, metode *contrasting group*, metode *contrasting group* dengan modifikasi *up and down*, metode *body of work*, dan metode *analytic judgment*; (4) metode berdasarkan pertimbangan kelompok peserta tes yang terdiri dari *Judgments about a Reference Group* dan *Judgments about Two Reference Groups*); dan (5) metode berdasarkan pada kompromi antara pertimbangan absolut dan normatif terdiri dari metode *Beuk* dan metode *Hofstee*.

Retnawati (2014, pp. 165–166) menggolongkan metode *standard setting* menjadi dua kelompok yaitu (1) metode berpusat pada butir/tes, metode ini menggunakan pendekatan klasik dan teori respon butir. Metode berpusat pada butir/tes yang cenderung menggunakan pendekatan klasik terdiri dari metode *Nedelsky*, penilaian pro-fesional, metode *Angoff*, dan metode *Ebel*. Sedangkan metode berpusat pada butir yang menggunakan pendekatan teori respon butir terdiri dari metode *Bookmark* dan metode pemetaan butir (*item mapping*); dan (2) metode yang berpusat pada siswa terdiri dari metode *Contrasting group* dan metode *Borderline*. Penggolongan metode *standard setting* didasarkan pada sudut pandang masing-masing ahli.

Penggolongan berbagai metode *standard setting* dapat mempermudah pemilihan metode *standard setting* yang sesuai dengan karakteristik, tujuan, dan situasi yang terjadi. Pemilihan metode *standard setting* yang tepat akan memberikan kesalahan (*error*) yang kecil. Sehingga penentuan *cut score* akan semakin tepat. Untuk itu sangat penting diperhatikan pemilihan metode *standard setting* yang akan digunakan. Metode *standard setting* yang digunakan akan sangat menentukan dalam menghasilkan *cut score* yang tepat dengan nilai *error* terkecil.

Cut score merupakan point penting dalam *standard setting*. Penentuan *cut score* bukanlah hal yang mudah (Nudell, 2008, p. 5). *Cut score* yang ditetapkan harus dapat mencerminkan ketercapaian kompetensi minimal yang harus dicapai peserta tes. Ketepatan penentuan *cut score* dalam *standard setting* ditentukan berdasarkan besar kecilnya *error*. Semakin besar nilai *error* maka semakin tidak tepat penentuan *cut score*. Sebaliknya, semakin kecil nilai *error* maka penentuan *cut score* semakin tepat. Selain itu, perlu diperhatikan juga bahwa penentuan *cut score* yang terlalu tinggi dapat menimbulkan kerugian bagi peserta tes. Hal ini dikarenakan *cut score* yang terlalu tinggi menyebabkan peserta tes yang seharusnya lulus menjadi tidak lulus. Sebaliknya jika *cut score* yang terlalu rendah memberikan keuntungan bagi peserta tes. Semakin rendah *cut score* yang ditentukan maka peserta tes yang seharusnya tidak lulus menjadi lulus. Besarnya *cut score* bisa dinaikkan ataupun diturunkan, tetapi penaikan dan penurunan *cut score* akan berdampak pada besar kecilnya nilai *error*.

Besar kecilnya *error* dari masing-masing *cut score* pada setiap metode *standard setting* menunjukkan ketepatan dari metode tersebut. Namun, penentuan *cut score* pada *standard setting* bukan hanya sebatas melihat besarnya *cut score* dan *errornya*,tetapi untuk menghasilkan *cut score* yang tepat dari metode *standard setting* sebaiknya dilakukan estimasi kesalahan pengukuran dari masing-masing metode *standard setting*.

Sejauh pengamatan peneliti, penelitian mengenai *standard setting* mayoritas melakukan perbandingan metode, tanpa mengestimasi kesalahan dari masing-masing metode. Dengan mengestimasi kesalahan dari setiap metode maka akan diperoleh *cut score* yang lebih akurat. Penelitian yang dilakukan oleh Prijowuntato, Mardapi, & Budiyono (2015, p. 176) merupakan salah satu penelitian yang membahas tentang estimasi kesalahan pengukuran *standard setting*. Penelitian tersebut menggunakan tiga metode *standard setting* yaitu metode *Angoff*, metode *Ebel*, dan

metode *Bookmark*. Estimasi kesalahan pengukuran menggunakan metode *Bootstrap*.

Penelitian serupa dilakukan oleh Yin & Sconing (2008, p. 182) tentang estimasi kesalahan *standard setting* dengan pendekatan *Generalizability Theory*, dimana metode yang digunakan terdiri dari metode *item rating* dan *Bookmark*. Pada penelitian tersebut, *Generalizability Theory* digunakan untuk mengestimasi kesalahan baku *cut score* yang dihasilkan oleh kedua metode *standard setting* yang digunakan yaitu *item rating* dan *Bookmark*. Selanjutnya, *Generalizability Theory* secara eksplisit menggabungkan beberapa sumber kesalahan di model pengukuran yang digunakan untuk mengestimasi kesalahan baku pada *cut score* dari masing-masing metode. Tujuan dalam penelitian tersebut ada tiga yaitu mengestimasi efek dari berbagai sumber kesalahan pada kedua prosedur; mengestimasi standar *error* pada *cut score* dari dua prosedur; dan mengestimasi efek perbedaan konseptualisasi yang berbeda dari seluruh bidang generalisasi untuk dua prosedur *standard setting*.

Berdasarkan latar belakang tersebut, maka estimasi kesalahan pengukuran *cut score* pada beberapa metode *standard setting* perl dilakukan. Metode *standard setting* yang digunakan dalam penelitian terdiri dari metode *Ebel*, metode *Bookmark*, dan metode *Contrasting group*. Metode *Ebel* dan metode *Bookmark* didasarkan pada tes/butir, dan metode *Contrasting group* didasarkan pada peserta tes (*examinee*). Ketiga metode ini memiliki prosedur yang berbeda dalam penentuan *cut score*. Perbedaan prosedur akan menghasilkan *cut score* dan nilai *error* yang berbeda. *Cut score* yang diperoleh dari ketiga metode ini kemudian diestimasi kesalahannya dengan pendekatan *Generalizability Theory* menggunakan program *eduG*.

Passing scores, cut scores, cut off scores, performance scores, achievement levels, mastery levels, prociency levels, thresholds levels, dan *standard* merupakan istilah–istilah dalam standar *setting* (Glass, 1978, p. 240). Istilah–istilah tersebut pada dasarnya memiliki makna yang sama yaitu bahwa *standard setting* merupakan suatu batas atau kriteria yang dijadikan dasar dalam suatu hal. *Standard setting* diartikan oleh Cizek (1996, p. 20) sebagai suatu proses dalam menentukan batas lulus *(cut score)*. Batas lulus tersebut merupakan batas bawah yang menentukan peserta didik dapat dikatakan kompeten atau tidak kompeten.

MacCann & Stanley (2006, p. 4) mendefinisikan *standard setting* sebagai kumpulan prosedur yang sistematik dalam mengidentifikasi batas lulus (*cut score*) yang diperlukan untuk menentukan tingkat kemahiran. Crocker & Algina, (1986, p. 410) menyebutkan *standard setting* sebagai kegiatan untuk menentukan skor batas lulus. Batas lulus tersebut menjadi kriteria dalam menentukan tingkat atau level prestasi seseorang. *Standard setting* merupakan suatu proses/ prosedur yang sistematis dalam menentukan batas lulus atau *cut score* untuk menyatakan tingkat prestasi.

Pengambilan keputusan berhubungan dengan prosedur–prosedur pengukuran. *Standard setting* merupakan prosedur pengukuran yang dapat digunakan dalam pengambilan keputusan. *Standard setting* adalah aturan yang dapat dipakai dalam pengambilan keputusan yang pentingdengan mempertimbangkan berbagai aspek.

Standard setting berperan penting dalam bidang pendidikan, tepatnya dalam menentukan batas kelulusan peserta didik. Penyelenggaraan ujian merupakan cara yang digunakan untuk mengevaluasi pembelajaran dengan melihat respon peserta didik terhadap tes yang dibuat. Ujian yang di Indonesia dikenal dengan Ujian Nasional memiliki batas kelulusan atau kriteria kelulusan dengan kata lain peserta tes tersebut dinyatakan kompeten terhadap suatu pelajaran atau materi jika hasil perolehannya melebihi kriteria yang ditentukan. Sebaliknya, seorang peserta tes dikatakan tidak lulus atau tidak kompeten terhadap suatu pelajaran atau materi apabila hasil perolehannya kurang dari kriteria yang ditentukan.

Penelitian ini menggunakan tiga metode *standard setting* yang terdiri dari dua metode berpusat pada tes dan satu metode berpusat pada peserta tes. Dua metode tersebut yaitu metode *Ebel* dan *Bookmark*. Me-

tode yang berpusat pada peserta tes yaitu metode *Contrasting group*.

Metode *Ebel* ini merupakan perbaikan dari metode *Angoff* dengan mempertimbangkan tingkat kesulitan butir dan relevansi butir. Tingkat kesulitan butir pada metode *Ebel* dibedakan menjadi tiga yaitu sulit, sedang, dan mudah. Sementara relevansi butir diperingkat menjadi empat yaitu *essential*, penting (*important*), dapat diterima (*acceptable*), dan dapat dipertanyakan (*questionable*) (Retnawati, 2014, p. 167). Prosedur ini menghasilkan tabel 3 x 4 dengan sejumlah butir tertentu yang diletakkan pada 12 sel kombinasi antara kesulitan butir dan relevansi butir (Alsmadi, 2007, p. 479).

Prosedur *Ebel* dapat dilakukan dengan penilai menentukan tingkat kesulitan butir (sulit, sedang, mudah) dan relevansi butir meliputi essensial, penting, dapat diterima, dan dapat dipertanyakan (Saunders, Ryan, & Huynh, 1980, p. 167); penilai mengisi butir-butir pada sel kombinasi tingkat kesulitan dan relevansi butir; penilai menentukan proporsi butir dalam masing-masing kategori bahwa *examinee* yang ada pada garis batas; penilai mengalikan jumlah butir dengan proporsi masing-masing. Hasil perkalian tiap kategori tersebut kemudian dijumlahkan. Hasil penjumlahan ini disebut *minimum passing score* (MPS); *cut score* diperoleh dengan merata–*rata minimum passing score* yang diusulkan oleh penilai.

Metode *Bookmark* dikembangkan untuk mengatasi keterbatasan yang berhubungan dengan *standard setting* yang terdahulu, khususnya untuk menentukan *cut score* ganda pada *single test* (Karantonis & Sireci, 2006, p. 6). Pengembangan metode ini mencakup beberapa prosedur seperti mengintegrasikan *selected response* dan *constructed response* ke dalam format butir, mudah diterapkan, dan mendasarkan pada IRT (Cizek & Bunch, 2007, p. 160). Pada metode *Bookmark*, butir-butir yang sudah diurutkan dengan menggunakan analisis IRT dikumpulkan dalam satu booklet yang disebut *Ordered Item Booklet* (OIB). Di samping itu, dalam metode *Bookmark* ditetapkan *Response Probability* (RP) sebesar 67% *likelihood*.

Adapun kelemahan implementasi prosedur *Bookmark* adalah komposisi soal dari yang termudah sampai paling sulit memungkinkan ada beberapa indikator dari kemampuan yang diujikan tidak termasuk dalam halaman *Bookmark* yang dipilih oleh panelis sebagai batas kemampuan siswa dalam menjawab. Disamping itu, kelemahan metode *Bookmark* adalah penilai kesulitan untuk memahami dan menggunakan kemungkinan jawaban (*Response Probability*).

Metode *Contrasting group* diperkenalkan oleh Berk pada tahun 1976. Berk menyarankan prosedur validasi kelompok yang diperluas. Prosedur kelompok yang menguasai digunakan untuk menentukan perbedaan *cut score* antara siswa-siswa yang terlatih dan tak terlatih, atau anatara siswa yang menguasai materi dan tidak menguasai materi (Cizek & Bunch, 2007, p. 106). Penentuan *cut score* pada metode ini menggunakan prosedur kelompok untuk membedakan kelompok master dan kelompok non-master. Dua distribusi kelompok tersebut kemudian digambar untuk menentukan titik potongnya.

Metode *Contrasting group* memiliki kelebihan dan kekurangan. Livingstone & Zieky (1982, p. 53) menganggap bahwa metode ini memiliki kemudahan dalam penerapannya dan memberikan hasil yang akurat. Metode ini didasarkan pada kondisi nyata peserta tes. Selain itu, tes dengan bentuk pilihan ganda cocok jika penentuan *cut score* dilakukan dengan metode ini, karena ahli akan menentukan kelompok master dan non-master dengan lebih mendasar. Sementara kekurangan metode ini adalah adanya kesulitan dalam memperoleh evaluasi yang sebanding untuk wilayah yang lebih luas, misal tingkat nasional.

Masalah dalam *standard setting* pada dasarnya sama dengan masalah yang dihadapi dalam pengukuran (Nichols, Twing, Mueller, & O'Malley, 2010, p. 19). Tidak tersedianya indikator dalam *standard setting* yang dapat digunakan untuk mengukur prestasi kelompok. Para panelis dalam *standard setting* diminta untuk membuat pertimbangan tentang kinerja kelompok siswa, sebagai

contoh siswa yang termasuk dalam kategori dasar, cukup, maupun maju.

Panelis yang terlibat dalam putaran dan skema dapat menyebabkan variabilitas dalam *cut score* (Yin & Sconing, 2008, p. 185). Kemungkinan besar bahwa perbedaan *cut score* dapat dihubungkan dengan perbedaan metode *standard setting* karena berbagai sumber yang disebutkan di atas dan atau perbedaan dalam prosedur. *Standard setting* termasuk prosedur pengukuran untuk menetapkan kemampuan siswa, maka variabilitas atau ketidakpastian dalam *cut score* yang dihasilkan dari proses *standard setting* perlu diperhatikan. *Standard error* dalam *cut score* seharusnya ikut dipertimbangkan. Namun demikian, sedikit penelitian yang mempertimbangkan *standard error* dalam *standard setting*.

Estimasi kesalahan pengukuran *cut score* dalam *standard setting* pada penelitian ini dilakukan dengan pendekatan *Generalizability Theory*. Pendekatan *generalizability theory* digunakan karena dalam metode *standard setting* ini digunakan panelis sebagai penentu *cut score*, dimana penggunaan panelis akan menimbulkan tingginya variabilitas nilai yang diberikan oleh panelis itu sendiri. Sehingga reliabilitas dari skor yang dihasilkan perlu diperhatikan. Penggunaan *generalizability theory* dikarenakan *G theory* memberikan berbagai model yang digunakan dalam menyelidiki kesalahan dalam metode *Ebel*, *Bookmark*, dan *Contrasting group*.

Analisis G-teori memiliki dua tahap yaitu generalisasi studi (G studi) dan *decision* studi (D studi). G studi dilakukan untuk menentukan seberapa baik skor dapat digunakan dalam beberapa situasi dan melibatkan perkiraan komponen varians yang mung kin akan digunakan dalam studi D. Studi D adalah penelitian yang dilakukan untuk menghitung koefisien reliabilitas dan SEs pengukuran dengan tujuan menentukan prosedur pengukuran yang paling efisien pada situasi tertentu. Pertimbangan studi D yang paling penting adalah spesifikasi dari generalisasi populasi dimana pembuat keputusan ingin menggeneralisasi skor dengan prosedur pengukuran tertentu.

Berdasarkan uraian yang telah disampaikan tersebut, maka penelitian ini bertujuan untuk mengetahui metode yang lebih akurat dalam mengestimasi kesalahan pengukuran *standard setting* pada metode *Ebel*, *Bookmark*, dan *Contrasting group*.

Metode

Penelitian ini termasuk jenis penelitian deskriptif kuantitatif. Sumber data dalam penelitian ini berupa respon siswa terhadap Ujian Nasional mata pelajaran Matematika pada jenjang SMP di Kabupaten Sumbawa tahun ajaran 2015/2016

Populasi dalam penelitian ini adalah lembar jawaban Ujian Nasional Matematika siswa dari 95 SMP Negeri di Kabupaten Sumbawa. Adapun sampel yang digunakan dalam penelitian ini merupakan respon jawaban Ujian Nasional Matematika siswa paket P0C5520 dari 12 SMP Negeri di Kabupaten Sumbawa dengan jumlah sampel sebanyak 352 siswa. Pengambilan sampel dalam penelitian ini menggunakan teknik *proportionate stratified random sampling*. Pengambilan sampel didasarkan pada klasifikasi sekolah dari kategori tinggi, sedang, dan rendah. Klasifikasi sekolah berdasarkan pada nilai Ujian Nasional tahun ajaran 2015/ 2016. Selain itu juga mempertimbangkan letak geografis sekolah baik itu di kota maupun di desa. Hal ini dilakukan agar sam-pel yang digunakan dalam penelitian dapat merepresentasikan keadaan sebenarnya di Kabupaten Sumbawa. Data tentang sampel penelitian disajikan pada Tabel 1.

Tabel 1. Daftar Sampel Penelitian

No	Nama Sekolah	Ketegori	Letak	Jumlah Sampel
1	SMP N 1 Sumbawa Besar	Tinggi	Kota	58
2	SMP N 1 Moyo Hilir	Tinggi	Desa	21
3	SMP N 1 Moyo Utara	Tinggi	Desa	25
4	SMP N 5 Moyo Hilir	Tinggi	Desa	10
5	SMP N 2 Labuhan Badas	Rendah	Desa	15
6	SMP N 2 Sumbawa Besar	Rendah	Kota	62
7	SMP N 3 Moyo Hilir	Rendah	Desa	10
8	SMP N 3 Sumbawa Besar	Rendah	Kota	32
9	SMP N 4 Labuhan Badas	Rendah	Desa	17
10	SMP N4 Sumbawa Besar	Rendah	Kota	7
11	SMP N 1 Unter Iwes	Rendah	Kota	34
12	SMP N 1 Labuhan Badas	Rendah	Desa	61
	Total			352

Panelis yang digunakan ditentukan dengan kuota sebanyak 12 orang guru matematika yang diambil berdasarkan kualitas sekolah (tinggi, sedang, dan rendah) dan letak geografisnya (kota dan desa) serta kriteria panelis. Adapun kriteria tersebut sebagai berikut: (1) ahli dalam bidang yang berhubungan dengan ujian; (2) terbiasa dengan metode–metode ujian; (3) telah mengajar matematika minimal 10 tahun; (4) mengajar matematika kelas 12 minimal selama 5 tahun; dan (5) lulusan dari program studi matematika atau pendidikan matematika.

Teknik yang digunakan untuk mengumpulkan data dalam penelitian ini adalah dokumentasi. Teknik dokumentasi digunakan untuk mengumpulkan respon jawaban siswa SMP di Kabupaten Sumbawa dalam menjawab Ujian Nasional mata pelajaran Matematika tahun ajaran 2015/2016. Selain itu, data dalam penelitian ini juga dikumpulkan melalui *Focus Group Discussion* (FGD). Adapun instrumen yang digunakan untuk mengumpulkan data dari FGD yaitu berupa lembar kerja panelis untuk metode *Ebel*, *Bookmark*, dan *Contrasting group*.

Teknik analisis data dalam penelitian ini terdiri dari tiga tahap yaitu tahap pertama atau persiapan terdiri dari penyiapan data, penggolongan SMP, dan pengujian karakteristik butir menggunakan program *Winstep*; tahap kedua yaitu *focus group discussion* (FGD) terdiri dari dua putaran; dan tahap ketiga yaitu mengestimasi kesalahan pengukuran dengan pendekatan *Generalizability Theory* menggunakan program *EduG*.

Hasil dan Pembahasan

Data respon Ujian Nasional Matematika siswa paket P0C5520 dari 12 SMP Negeri di Kabupaten Sumbawa dengan jumlah sampel sebanyak 352 siswa dengan soal terdiri dari 40 butir. Selanjutnya dianalisis menggunakan program *Winstep* untuk menghitung tingkat kesulitan butir. Tingkat kesulitan butir pada *output Winsteps* dapat dilihat pada *Table Measure*. Hasil analisis *Winsteps* tampak pada Tabel 2.

Tabel 2. Tingkat Kesulitan Butir

No	Kriteria	Nomor Butir	Jumlah
1	Mudah	1, 10, 16, 39	4
2	Sedang	4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 37, 38, 40	30
3	Sulit	2, 3, 17, 19, 23, 36	6
Jumlah			40

Berdasarkan Tabel 2 diketahui bahwa sebanyak 4 butir (10%) tergolong butir yang mudah, 30 butir (75%) termasuk butir yang tergolong sedang atau yang biasa disebut sebagai butir yang baik, dan butir yang termasuk dalam kategori butir sulit berjumlah 6 butir (15%).

Tahap kedua yaitu FGD untuk menentukan *cut score* dari metode *Ebel*, *Bookmark*, dan *Contrasting group*. Tahap ini dilakukan dengan dua putaran. Masing-masing metode memiliki dua *cut score*. Berikut *cut score* tiap metode disajikan pada Tabel 3.

Tabel 3. *Cut Score* Putaran Pertama dan Kedua

Putaran	Metode *Standard Setting*		
	Ebel	Bookmark	Contrasting Group
Pertama	64,579	64,434	50
Kedua	67,813	65,455	47,5

Ketiga metode menghasilkan *cut score* yang berbeda. Perbedaan ini terjadi dikarenakan prosedur penentuan *cut score* masing-masing metode sangat berbeda. Metode *Ebel* memiliki *cut score* tertinggi pada kedua putaran yaitu 64,579 dan 67,813 dengan jumlah siswa yang lulus sebanyak 128 siswa dan mampu menguasai minimal 26 deskriptor. *Cut score* metode *Bookmark* putaran 1 dan 2 berturut 64,434 dan 65,455. Jumlah siswa yang lulus pada *cut score* metode *Bookmark* sebanyak 128 siswa dengan deskriptor yang harus dikuasai sebanyak 25 deskriptor. Metode *Contrasting group* memiiki *cut score* paling rendah pada kedua putaran dengan jumlah siswa yang lulus sebanyak 297 siswa. Adapun siswa dikatakan lulus jika mampu menguasai 19 deskriptor.

Rendahnya kelulusan siswa tersebut dapat disebabkan oleh kekurang telitian peserta didik dalam mengerjakan soal. Walaupun soal yang diujikan termasuk dalam kriteria sedang, namun apabila peserta didik tidak teliti dalam mengerjakan maka hasil pengerjaan tetap salah. Padahal berdasarkan pendapat panelis soal-soal yang diujikan termasuk dalam kategori esensial dan penting, yang mana berarti materi-materi yang berkaitan dengan soal tersebut harusnya dapat dipahami dan dimengerti oleh peserta didik.

Berdasarkan hasil penelitian ini penyebab tingginya *cut score* pada metode *Ebel* dikarenakan masih kurangnya pemahaman guru tentang relevansi butir dengan kategori esensial, penting, dapat diterima, dan dapat dipertanyakan serta tingkat kesulitan butir. Selain itu, perkiraan proporsi jawaban benar untuk masing-masing kategori juga menjadi penyebab tingginya *cut score*. Kurangnya pemahaman guru akan hal-hal penting pada metode *Ebel* membuat guru dengan bebas meletakkan soal pada keempat kategori dengan proporsi jawaban benar yang tinggi sesuai dengan kemampuan guru dalam menginterpretasi soal.

Prosedur *Bookmark* merupakan prosedur yang dirasa mudah bagi panelis dalam menentapkan *cut score*. Hal ini dikarenakan data yang digunakan untuk menentukan *cut score* sudah disiapkan terlebih dahulu. Sehingga dapat mempermudah panelis dalam menentukan halaman bookmark yang dianggap sesuai dengan kemampuan peserta didik menjawab benar.

Berbeda dengan metode *Ebel* dan *Bookmark*, pada metode metode *Contrasting group* putaran pertama dan kedua diperoleh *cut score* yang lebih rendah dari kedua metode tersebut. Karakteristik parameter butir soal secara tidak langsung akan mempengaruhi hasil perhitungan *cut score* yang diperoleh dengan menggunakan metode *Contrasting Group*. Jika butir-butir soal yang terdapat dalam perangkat tes memiliki tingkat kesulitan rendah maka butir soal akan dijawab dengan benar sehingga skor peserta tes menjadi tinggi. Sebaiknya jika butir-butir soal yang terdapat dalam perangkat tes memiliki tingkat kesulitan tinggi maka butir soal akan susah dijawab dengan benar dan menyebabkan skor peserta tes menjadi rendah. Distribusi skor inilah yang mempengaruhi *cut score* yang diperoleh.

Setelah diperoleh *cut score* pada ketiga metode *standard setting*, selanjutnya dilakukan tahap ketiga. Tahap ketiga yaitu mengestimasi kesalahan pengukuran dengan pendekatan *generalizability theory*. Pendekatan ini dilakukan dengan dua tahap yaitu G studi dan D studi.

G studi dilakukan untuk menentukan seberapa baik skor dapat digunakan dan melibatkan perkiraan varian komponen yang akan digunakan dalam studi D. D studi dilakukan untuk menghitung koefisien reliabilitas dan SEs pengukuran dengan tujuan menentukan prosedur pengukuran yang paling efisien. Analisis pada pendekatan ini dilakukan dengan program *EduG*. Hasil analisis dapat dilihat pada Tabel 4.

Tabel 4. Estimasi Varians Komponen dengan Desain G studi

VC	Ebel		Bookmark		Contrasting group	
	Estimasi	%	Estimasi	%	Estimasi	%
	31,41	76,8	5,35	77,4	0,00	0,0
	5,08	13,6	0,39	2,4	3,12	100
	1,79	9,6	1,60	20,1	0,00	0,0

Hasil menunjukkan bahwa estimasi varians komponen putaran terbesar pada metode *Ebel*, *Contrasting group*, dan *Bookmark*. Namun dilihat dari proporsinya varian komponen putaran pada metode *Contrasting group* lebih tinggi dibanding metode *Ebel* dan *Bookmark*. Tingginya proporsi varians komponen putaran pada metode *Contrasting group* dikarenakan pada metode ini hanya terdapat dua *cut score* yaitu putaran satu dan dua. Dimana *cut score* setiap panelis untuk masing-masing putaran itu sama. Sehingga menyebabkan variabilitas *cut score* hanya tinggi pada putaran. Variabilitas inilah yang akan berpengaruh pada koefisien reliabilitas.

Setiap G studi akan diperoleh nilai koefisien G *relative* dan G *absolute* pada masing-masing metode. Koefisien G diguna-

kan untuk menentukan besarnya koefisien reliabilitas setiap metode atau dengan kata lain koefisien G disebut sebagai koefisien reliabilitas. Adapun koefisien G dapat dilihat pada Tabel 5.

Tabel 5. Koefisien G pada Metode *Ebel*, *Bookmark* dan *Contrasting group*

Koefisien	Ebel	Bookmark	Contrasting group
G relative	0,94	0,59	1,00
G Absolute	0,65	0,23	1,00

Tabel 5 menunjukkan bahwa koefisien G tertinggi dari ketiga metode tersebut yaitu diperoleh pada metode *Contrasting group* sebesar 1,00. Koefisien G terendah diperoleh oleh metode *Bookmark*. Hal ini berarti bahwa metode *Contrasting group* memiliki koefisien reliabilitas tertinggi atau dengan kata lain metode *Contrasting group* lebih reliabel daripada metode *Ebel* dan *Bookmark*. Selanjutnya dilakukan D studi untuk melihat reliabilitas dari masing-masing optimalisasi pada setiap metode.

Tahap kedua pada pendekatan *Generalizability Theory* yaitu D studi. D studi dilakukan untuk menghitung koefisien reliabilitas dan *standard error* pada masing-masing metode. *Standard error* yang digunakan berdasarkan hasil analisis dengan eduG dapat dilihat pada *absolute standard error of measurement* (SEM). Masing-masing metode *standard setting* terdiri dari lima SEM yang berbeda. Hasil SEM pada *cut score* masing-masing metode dapat dilihat pada Tabel 6.

Tabel 6. *Standard Error of Measurement* pada *Cut Score*

SEM	Ebel	Bookmark	Contrasting Group
		(1)	
	1,158	0,803	0,000
		(2)	
	1,113	0,772	0,000
		(3)	
	1,072	0,744	0,000
		(4)	
	1,036	0,718	0,000
		(5)	
	1,036	0,696	0,000

Berdasarkan hasil tersebut terlihat bahwa semakin banyak jumlah panelis yang digunakan dengan jumlah putaran yang sama maka nilai SEM semakin rendah. Nilai SEM tertinggi pada metode *Ebel* yang mencapai 1,158 dengan jumlah panelis sebanyak 12 orang. Nilai SEM terendah pada metode *Contrasting group* sebesar 0,000 untuk semua pola optimalisasi.

Estimasi varians komponen pada ketiga metode *standard setting* menunjukkan metode *Contrasting group* memiliki nilai estimasi terendah pada efek panelis dan interaksi panelis dan putaran. Estimasi varian komponen pada efek dan interaksi panelis dan putaran tertinggi pada metode *Ebel*. Estimasi VC tertinggi menunjukkan tingginya variabilitas nilai yang diberikan panelis dan sumber lain, seperti tingkat kesulitan butir soal dan tingkat penguasaan peserta didik yang membutuhkan *judgment* panelis. Hal ini senada dengan yang disampaikan oleh Lee & Lewis (2008, p. 614).

Hasil varian komponen yang didasarkan pada G studi menjadi dasar dilakukannya D studi. Hasil D studi digunakan untuk menentukan koefisien reliabilitas, SEM dan keterkaitan *universe* untuk bisa digeneralisasi. Perbandingan nilai SEM untuk menentukan metode yang paling akurat dalam menentukan *cut score*. Nilai SEM terkecil diantara ketiga metode merupakan metode paling akurat.

Seperti yang diharapkan, *universe* pada bentuk random menghasilkan SE yang lebih besar karena *universe* memiliki definisi yang lebih luas untuk digeneralisasi daripada *universe* bentuk *fixed*. *Universe* bentuk *fixed* lebih dibatasi, akibatnya menghasilkan SE yang lebih kecil. Secara umum, SE *cut score* metode *Contrasting group* relatif lebih kecil daripada SE *cut score* metode *Ebel* dan *Bookmark*.

SEM pada ketiga metode dengan jumlah panelis dua belas dan putaran sebanyak dua kali. Hal ini menunjukkan bahwa nilai SEM pada metode *Contrasting group* lebih kecil dari metode *Ebel* dan metode *Bookmark*. Berarti bahwa metode *Contrasting group* lebih akurat daripada dua metode lainnya.

Rendahnya nilai SEM pada metode *Contrasting group* dikarenakan prosedur metode ini tidak membutuhkan penilaian dari setiap panelis, yang mana panelis melakukan diskusi untuk menentukan kelompok master dan nonmaster berdasarkan pengetahuan dan pengalaman panelis selama mengajar. Penentuan kelompok master dan nonmaster dilakukan hanya pada pembagian sekolah. Selanjutnya para panelis secara bersama menentukan jumlah peserta didik yang mampu mengerjakan butir dengan benar berdasarkan interval yang digunakan dalam penelitian ini. Kemudian terbentuklah distribusi frekuensi yang dijadikan sebagai penentu *cut score*. Sehingga hanya dihasilkan satu *cut score* dari dua belas panelis tanpa adanya istilah merata-ratakan *cut score* panelis.

Prosedur pada metode *Contrasting group* sangat berbeda dengan dua metode lainnya. Dimana metode *Ebel* dan *Bookmark* menghasilkan dua belas *cut score* yang selanjutnya dirata-ratakan untuk dijadikan sebagai *cut score* metode tersebut. Selain itu, metode *Contrasting group* merupakan metode yang paling mudah untuk diterapakan dibandingkan metode *Ebel* dan *Bookmark*. Sehingga, menyebabkan metode *Ebel* dan *Bookmark* memiliki variabilitas nilai yang tinggi.

Selain itu, koefisien G dari ketiga metode menunjukkan bahwa metode *Contrasting group* memiliki koefisien G yang lebih besar dari metode *Ebel* dan *Bookmark*. Hal ini berarti bahwa metode *Contrasting group* lebih reliabel dibandingkan metode *Ebel* dan *Bookmark*. Dikarenakan variabilitas *cut score* pada metode *Contrasting group* rendah.

Simpulan

Berdasarkan hasil penelitian yang telah dilakukan terkait dengan perbandingan estimasi kesalahan pengukuran *standard setting* pada penilaian kompetensi Matematika SMP di Kabupaten Sumbawa dapat disimpulkan bahwa *cut score* mata pelajaran Matematika Jenjang SMP di Kabupaten Sumbawa yang dihasilkan dengan menggunakan metode *Ebel* pada putaran pertama adalah 64,579 mengalami kenaikan pada putaran kedua menjadi 67,813. Pada metode ini siswa dikatakan lulus jika mampu menguasai 26 deskriptor. Selanjutnya pada metode *Bookmark* diperoleh *cut score* putaran pertama sebesar 64,434 mengalami penurunan pada putaran kedua menjadi 65,455, dimana siswa dikatakan lulus jika menguasai 25 deskriptor.

Cut score mata pelajaran Matematika jenjang SMP di Kabupaten Sumbawa yang dihasilkan dengan menggunakan metode *Contrasting group* putaran 1 sebesar 50,00 mengalami penuruna pada putaran kedua menjadi47,5. Terdapat 19 deskriptor kemampuan Matematika yang harus dikuasai oleh siswa agar dapat lulus berdasarkan metode *Contrasting group*. Metode *Contrasting group* merupakan metode yang paling akurat untuk mengestimasi kesalahan pengukuran dibandingkan metode *Ebel* dan *Bookmark*. Hal ini dikarenakan variabilitas *cut score* pada metode *Contrasting group* rendah, memiliki nilai SEM yang rendah, dan koefisien G yang tinggi.

Saran yang dapat diberikan berdasarkan simpulan penelitian ini adalah sebagai berikut. Bagi Dinas Pendidikan Kabupaten Sumbawa, perlu adanya pelatihan mengenai penentuan *cut score* bagi kelompok guru mata pelajaran sehingga guru memiliki tambahan pengetahuan yang lebih dan akhirnya mampu menerapkan metode-metode tersebut dalam menentukan *cut score*; bagi peneliti berikutnya, perlu dilakukan penelitian lebih lanjut terkait dengan *cut score* tentang penguasaan materi matematika di sekolah menengah pertama; dan bagi peneliti berikutnya, perlu dilakukan penelitian lebih lanjut terkait dengan *cut score* tentang penguasaan materi matematika di sekolah menengah pertama.

Daftar Pustaka

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

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

https://doi.org/10.3102/00346543056001137

Cizek, G. J. (1996). An NCME instructional module on: setting passing scores. *Educational Measurement: Issues and Practice*, *15*(2), 20–31. https://doi.org/10.1111/j.1745-3992.1996.tb00809.x

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: a guide to establishing and evaluating performance standards for tests*. California: Sage Publication, Inc.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: CBS College Publishing.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, *15*(4), 237–261. https://doi.org/10.1111/j.1745-3984.1978.tb00072.x

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: a literature review. *Educational Measurement: Issues and Practice*, *25*(1), 4–12. https://doi.org/10.1111/j.1745-3992.2006.00047.x

Lee, G., & Lewis, D. M. (2008). A Generalizability theory approach to standard error estimates for bookmark standard settings. *Educational and Psychological Measurement*, *68*(4), 603–620. https://doi.org/10.1177/0013164407312603

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.

MacCann, R. G., & Stanley, G. (2006). The use of rasch modeling to improve standard setting. *Practical Assessment, Research & Evaluation*, *11*(2), 1 – 17.

Menteri Pendidikan dan Kebudayaan Republik Indonesia. Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 5 Tahun 2015 tentang Kriteria Kelulusan Peserta Didik, Penyelenggaraan Ujian Nasional, dan Penyelenggaraan Ujian Sekolah/Madrasah/Pendidikan Kesetaraan Pada Smp/Mts atau yang Sederajat d (2015).

Nichols, P., Twing, J., Mueller, C. D., & O'Malley, K. (2010). Standard-setting methods as measurement processes. *Educational Measurement: Issues and Practice*, *29*(1), 14–24. https://doi.org/10.1111/j.1745-3992.2009.00166.x

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

Prijowuntato, S. W., Mardapi, D., & Budiyono, B. (2015). Perbandingan estimasi kesalahan pengukuran standard setting dalam penilaian kompetensi akuntansi SMK. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *19*(2). Retrieved from https://journal.uny.ac.id/index.php/jpep/article/view/5578

Retnawati, H. (2014). *Teori respon butir dan penerapannya*. Yogyakarta: Nuha Medika.

Saunders, J. C., Ryan, J. P., & Huynh, H. (1980). *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.

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), 182–197.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: a manual for setting standards of performance on educational and occupational tests*. Princeton, New Jersey: Educational Testing Service.