AKURASI METODE KALIBRASI *FIXED PARAMETER*: STUDI PADA PERANGKAT UJIAN NASIONAL MATA PELAJARAN MATEMATIKA

Dina Huriaty, Djemari Mardapi
SMP Negeri 1 Kertak Hanyar Kalimantan Selatan, Universitas Negeri Yogyakarta
dina_rty@yahoo.co.id, djemarimardapi@yahoo.com

Abstrak

Penelitian ini bertujuan untuk (1) mengidentifikasi karakteristik butir-butir tes pada perangkat soal ujian nasional mata pelajaran Matematika tingkat SMP tahun pelajaran 2009/2010 yang dikalibrasi dengan metode kalibrasi *fixed parameter*, dan (2) mengetahui metode kalibrasi *fixed parameter* yang paling akurat di antara metode NWU-OEM (*no prior weights updating and one expectation-maximization cycle*), NWU-MEM (*no prior weights updating and multiple expectation-maximization cycles*), OWU-OEM (*one prior weights updating and one expectation-maximization cycle*), OWU-MEM (*one prior weights updating and multiple expectation-maximization cycles*), dan MWU-MEM (*multiple weights updating and multiple expectation-maximization cycles*). Penelitian ini menggunakan pendekatan kuantitatif deskriptif. Subjek penelitian adalah data respons ujian nasional mata pelajaran Matematika tingkat SMP tahun pelajaran 2009/2010 dari provinsi DI Yogyakarta. Kriteria akurasi metode adalah nilai fungsi informasi tes dan kesalahan pengukuran. Hasil penelitian adalah sebagai berikut. (1) Statistik parameter butir-butir tes pada perangkat ujian nasional mata pelajaran Matematika tingkat SMP tahun pelajaran 2009/2010 menunjukkan rerata indeks daya beda butir berada pada interval [1,07 sampai 1,14], rerata indeks kesukaran butir [-0,35 sampai -0,20], dan rerata *pseudo guessing* < 0,25. Nilai theta-nilai kemampuan-pada posisi fungsi informasi butir menjadi maksimal menunjukkan grafik fungsi kelima metode kalibrasi *fixed-parameter* hampir berimpit. (2) Metode OWU-OEM merupakan metode yang paling akurat dalam mengestimasi parameter butir pada perangkat tes ujian nasional mata pelajaran Matematika tahun pelajaran 2009/2010.

Kata kunci: *akurasi, kalibrasi, fixed parameter, algoritma, Expectation-Maximization*

THE ACCURACY OF THE FIXED PARAMETER CALIBRATION METHOD: STUDY OF MATHEMATICS NATIONAL EXAMINATION TEST

Dina Huriaty, Djemari Mardapi
SMP Negeri 1 Kertak Hanyar Kal-Sel, Universitas Negeri Yogyakarta
dina_rty@yahoo.co.id, djemarimardapi@yahoo.com

Abstract

This study aimed to: (1) identify the characteristics of the test items on the mathematics test of the national examination which are calibrated with the fixed parameter calibration methods, and (2) reveal the most accurate fixed parameter calibration methods among NWU-OEM (no prior weights updating and one expectation-maximization cycle), NWU-MEM (no prior weights updating and multiple expectation-maximization cycles), OWU-OEM (one prior weights updating and one expectation-maximization cycle), OWU-MEM (one prior weights updating and multiple expectation-maximization cycles), and MWU-MEM (multiple weights updating and multiple expectation-maximization cycles) methods. This study used descriptive quantitative approach. The subject is the testee' responses to the mathematics national examination in junior high school in 2009/2010. The criteria of the accuracy methods are TIF and SEM. The research results are as follows. (1) Item of statistical parameter on Mathematics national examination test in 2009/2010 showed the average of item discrimination on the interval [1.07, 1.14], the average of item difficulty on the interval [-0.35, -0.20], and the average of *pseudo guessing* is c < 0.25. Theta - ability - score where the item information function maximalist showed the function of five *fixed-parameter* calibration methods almost coincides. (2) OEM-OWU method is the most accurate in estimating the parameters on mathematics national examination test in 2009/2010.

Keywords: *Accuracy, Calibration, Fixed Parameter, Algorithm, Expectation-Maximization*

Pendahuluan

Usaha pemerintah dalam rangka menjaga, mengendalikan, dan meningkatkan kualitas pendidikan di Indonesia dilakukan secara terus menerus dan berkesinambungan melalui suatu sistem penilaian pendidikan yang diatur dalam standar penilaian pendidikan. Standar penilaian pendidikan merupakan standar nasional pendidikan yang berkaitan dengan mekanisme, prosedur, dan instrumen penilaian hasil belajar siswa pada tingkat pendidikan dasar dan menengah. Kegiatan pengukuran pencapaian kompetensi pada siswa, baik pada tingkat pendidikan dasar dan menengah, dalam rangka pencapaian standar nasional pendidikan dilakukan melalui ujian nasional.

Ujian nasional dilaksanakan berdasarkan Peraturan Pemerintah No. 19 Tahun 2005 dan perubahannya, Peraturan Pemerintah No. 32 Tahun 2013 tentang standar nasional pendidikan. Peraturan Pemerintah ini menjelaskan bahwa ujian nasional bertujuan untuk menilai pencapaian kompetensi lulusan secara nasional pada sejumlah mata pelajaran tertentu. Selanjutnya dijelaskan bahwa hasil ujian nasional digunakan sebagai salah satu pertimbangan untuk pemetaan mutu satuan pendidikan, dasar seleksi masuk jenjang pendidikan berikutnya, penentuan kelulusan peserta didik dari program dan/atau satuan pendidikan, dan dasar pembinaan dan pemberian bantuan kepada satuan pendidikan dalam upaya peningkatan mutu pendidikan. Ujian nasional termasuk dalam kategori tes prestasi belajar yang pelaksanaannya dilakukan setiap akhir jenjang pendidikan.

Pelaksanaan ujian nasional di Indonesia menggunakan sejumlah perangkat soal yang dibuat setara. Hal ini membutuhkan sejumlah besar butir-butir soal. Butir-butir soal yang digunakan secara teoretik telah melalui tahapan-tahapan pengembangan tes hasil belajar. Proses pengembangan tes hasil belajar menurut Depdiknas (2000), meliputi tahapan: (1) penentuan tujuan tes; (2) penyusunan kisi-kisi tes; (3) penulisan soal; (4) penelaahan soal (*review dan revisi* soal); (5) uji coba soal menjadi instrumen tes; (6) perakitan soal menjadi instrumen tes; (7) penyajian tes; (8) pensekoran; (9) pelaporan hasil tes, dan (10) pemanfaatan hasil tes. Menurut Hambleton & Jones (1993), proses pengembangan tes dapat dilakukan melalui tahapan berikut: (1) penyiapan spesifikasi tes; (2) penyiapan pool soal tes; (3) pengujian soal di lapangan; (4) revisi soal tes; (5) pengembangan tes; (6) pilot testing (pengujicobaan instrumen hasil revisi); (7) pengembangan tes final; (8) pelaksanaan pengujian; (9) teknik analisis; (10) penyiapan petunjuk pelaksanaan; dan (11) pencetakan dan distribusi tes dan manual.

Pada tahapan uji coba termasuk di dalamnya adalah proses analisis butir dan penentuan parameter butir atau yang disebut kalibrasi butir. Kalibrasi butir adalah proses estimasi untuk menentukan parameter-parameter butir. Menurut *Standards for Educational and Psychological Testing* (1999, p.172), kalibrasi dalam teori respons butir merupakan proses estimasi parameter-parameter butir. Fungsi respons suatu butir memuat dua parameter, yaitu parameter butir dan parameter orang. Kalibrasi butir dilakukan untuk mengestimasi parameter butir berdasarkan model teori respons butir (*Item Response Theory*).

Teori respons butir (TRB) yang juga disebut teori tes modern menjadi kerangka kerja statistik untuk menyelesaikan masalah pengukuran, seperti pengembangan tes. Manfaat utama penggunaan pendekatan teori respons butir adalah (1) parameter butir bersifat invarian, dimana fungsi/kurva respons butir tidak berubah; dan (2) seleksi butir-butir berdasarkan atas banyaknya informasi butir dan informasi tes (Hambleton, Swaminathan & Rogers, 1991, p.100). Manfaat utama penggunaan pendekatan TRB dalam konstruksi tes, antara lain (1) proses kalibrasi butir dengan *sample free*, dan (2) pencapaian pengukuran orang dengan *item free* (Green, Yen & Burket, 1989). Analisis tes dengan pendekatan IRT ini menempatkan parameter soal bersifat invarian, artinya fungsi atau kurva dari respons butir tidak berubah, walaupun kelompok peserta yang menjawab butir yang sama itu berubah-

ubah. Selain itu, tingkat kesulitan butir dan kemampuan peserta tes diukur dengan skala yang sama sehingga dimungkinkan untuk menyeleksi butir-butir tertentu berkenaan dengan skala kemampuan.

Pada tes-tes standar, diasumsikan bahwa parameter butir telah diketahui. Hal ini dikarenakan kalibrasi butir dilakukan selama proses standarisasi tes. Pada tes yang baru diujikan, parameter-parameter butir akan diestimasi dari data yang diperoleh. Selain mengestimasi parameter butir, parameter kemampuan juga diestimasi. Pada berbagai situasi diperlukan lebih dari satu perangkat tes yang diujikan (Wright & Stone, 1979, p.98). Pada situasi ini kadangkala butir bersama (*common items*) ditempatkan pada perangkat tes tersebut. Sebagai contoh, jika ingin mengestimasi 60 butir sedangkan hanya diperlukan 40 butir yang akan diujikan pada setiap peserta, maka 60 butir dapat dibagi menjadi tiga set tes, yaitu A, B, dan C. Dua perangkat tes dapat dibentuk, perangkat pertama memuat A dan B, sedangkan perangkat kedua memuat A dan C. Kedua perangkat diujikan pada dua kelompok peserta tes yang berbeda. Respons kedua kelompok terhadap butir-butir pada tes A dapat digunakan untuk membuat skala umum terhadap parameter semua butir yang diestimasi.

Sejumlah *common items* digunakan sebagai pengait untuk menghitung estimasi parameter butir dari dua buah perangkat pada skala yang sama (Lee & Ban, 2010). Penempatan skala umum pada dua atau lebih tes memungkinkan untuk membandingkan tingkat kesukaran tes dan juga menjadi dasar dalam pengembangan bank soal. Menurut Hambleton, Swaminathan, & Rogers (1991, p.128) ada empat desain penyetaraan yang dapat digunakan untuk penskalaan parameter butir yaitu (1) *Single-Group Design*, (2) *Equivalent-Group Design*, (3) *Anchor-Test Design*, dan (4) *Common-Person Design*, sedangkan menurut Kolen & Brennan (1995, pp.13-19) ada tiga desain pengumpulan data, yaitu: (1) *Random-Group Design,* (2) *Single-Group Design*, dan (3) *Common-Items Nonequivalent-Group Design*. Pada desain *random group*, dua buah tes yang akan disetarakan diujikan pada kelompok uji yang sama. Desain ini sederhana, tetapi tidak praktis karena membutuhkan waktu yang lama dalam pelaksanaannya. Desain yang lebih praktis dalam pelaksanaannya dan tidak ada efek akibat mengulang dan kelelahan adalah *Equivalent-Group Design*. Pada desain ini dua buah tes yang akan disetarakan/dikaitkan, diujikan pada dua kelompok yang ekuivalen. Pemilihan kelompok dilakukan secara acak. Pada *Anchor-Test Design*, tes-tes dengan butir bersama diberikan kepada dua kelompok uji. Setiap tes mempunyai sejumlah butir bersama. Desain ini mudah dilaksanakan dan seringkali digunakan. Jika pemilihan *common-item* dilakukan dengan tepat maka desain ini dapat menjadi alternatif untuk mengatasi masalah pada desain *Single-Group* dan *Equi-valent-Group*. Pada desain *Common-Person*, dua buah tes yang dikaitkan diberikan pada dua kelompok dengan kelompok bersama mendapatkan kedua tes tersebut. Kelemahan dari desain ini adalah akan terjadi efek kelelahan pada kelompok bersama, karena kelompok tersebut akan mendapatkan tes dalam jumlah yang lebih banyak dibandingkan peserta yang tidak termasuk dalam kelompok bersama.

Desain yang digunakan pada perangkat soal ujian nasional tahun pelajaran 2009/2010 adalah dua kelompok peserta tes yang berbeda, masing-masing memperoleh naskah tes yang berbeda juga, dan pada setiap naskah tes berisi kumpulan *anchor-items (common items).* Untuk keperluan ini, maka desain *anchor-test* dipilih untuk keperluan yang praktis. Pada *anchor-test design* setiap perangkat tes yang dibandingkan menggunakan soal-soal bersama atau *common items* sebagai dasar untuk membandingkan perbedaan kemampuan dua kelompok yang mengerjakan perangkat tes yang berbeda. Ujian nasional tahun pelajaran 2009/2010 menggunakan dua tes yang berbeda tetapi berasal dari kisi-kisi tes yang sama, dengan *common items* yang diujikan kepada dua kelompok siswa, sehingga parameter butir yang diestimasi dari dua kelompok tersebut tidak dijamin berada pada skala yang sama.

Cara untuk menempatkan parameter estimasi dari dua kelompok yang terpisah kedalam skala yang sama, dapat dilakukan dengan menghitung parameter estimasi untuk setiap kelompok dan kemudian mengubah skala dengan menggunakan *common-items*. Cara yang lain adalah dengan menetapkan parameter *common-items* kemudian mengestimasi parameter butir yang bukan *common-items* secara bersama-sama, sehingga berada pada skala yang sama dengan *common-items*.

Common items mempunyai peranan penting dalam kalibrasi butir, karena itu ketika desain *anchor-test* digunakan, hendaknya memperhatikan sifat dan karakteristik dari *common items* dan penggunaan skornya. *Common items* seharusnya 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. Jumlah *common items* disarankan minimal 20% dari panjang tes untuk model tes yang diskor secara dikotomus (Kolen & Brennan, 1995, p.248). Secara umum desain *common test* memerlukan ukuran sampel yang besar. Desain *anchor-test* ini sering digunakan, karena dua kelompok peserta tes yang dibutuhkan tidak harus sama kemampuannya dan berasal dari populasi yang berbeda.

Ada tiga cara kalibrasi yaitu kalibrasi terpisah (*separate calibration)*, bersama-sama (*concurrent calibration*), maupun tetap (*fixed calibration*). Pada beberapa penelitian (Li, et al, 1997; Ban, et al, 2001, Taehoon & Petersen, 2009). Kalibrasi *fixed parameter* disebut sebagai *fixed item parameter calibration* dan *fixed abc*. Li, et al. (1997) menyatakan bahwa kalibrasi *Fixed ABC* menghasilkan hasil penyetaraan yang lebih stabil, terutama untuk parameter *c* dan θ pada penyetaraan horizontal. Hasil kalibrasi *fixed parameter* terhadap estimasi θ dan kalibrasi terpisah secara umum konsisten. Hasil simulasi menunjukkan bahwa kalibrasi *fixed parameter* dan kalibrasi bersama-sama menghasilkan estimasi parameter butir dan kemampuan yang sangat akurat dan stabil. Hasil penelitian tersebut juga menunjukkan bahwa kalibrasi *fixed parameter* menghasilkan estimasi yang lebih akurat dan stabil dibandingkan kalibrasi bersama-sama pada tes yang relatif mudah. Menurut Taehoon & Petersen (2009), kalibrasi *fixed parameter* menunjukkan hasil yang konsisten dibandingkan dua metode kalibrasi yang lainnya, yaitu kalibrasi bersama-sama (*concurrent calibration*) dan kalibrasi terpisah (*separate calibration*).

Kalibrasi *fixed parameter* disebut dalam penelitiannya Kim (2006). Kim (2006) membedakan metode kalibrasi *fixed parameter* berdasarkan algoritma *expectation-maximization*, yaitu (1) metode NWU-OEM (*no prior weights updating and one expectation-maximization cycle*), (2) NWU-MEM (*no prior weights updating and multiple expectation-maximization cycles*), (3) OWU-OEM (*one prior weights updating and one expectation-maximization cycle*), (4) OWU-MEM (*one prior weights updating and multiple expectation-maximization cycles*), dan (5) MWU-MEM (*multiple weights updating and multiple expectation-maximization cycles*). Kelima metode ini merupakan metode estimasi kebolehjadian maksimum marjinal yang didasari pada algoritma *expectation-maximization*. Metode kalibrasi *fixed parameter* ini dibedakan berdasarkan banyaknya kemampuan prior yang diperbaharui atau di-*update* selama siklus EM dan banyaknya siklus EM yang digunakan.

Hasil penelitian Kim (2006) menunjukkan bahwa metode MWU-MEM menghasilkan performa yang akurat pada ketiga distribusi kemampuan θ untuk kelompok target, sedangkan pada distribusi N(0,1), metode NWU-MEM dan metode OWU-MEM yang menunjukkan performa yang akurat dalam mengestimasi parameter.

Metode kalibrasi *fixed parameter* menghasilkan skala bersama dengan cara memperbaiki dan mentransformasi estimasi parameter butir untuk *common items*. Kim (2006) membedakan metode kalibrasi *fixed parameter* berdasarkan sejumlah *prior* yang terbaru. Pada metode *no prior update*, distribusi kemampuan *prior* diasumsikan berdistribusi normal N(0,1) dan tidak diperbaharui selama siklus *Expectation-Maximization (EM)*. Pada metode *iterative atau multiple prior update*,

distribusi kemampuan *prior* secara iterasi diperbaharui melalui program komputer yang dijalankan berkali-kali untuk membangkitkan *fixed parameter*. Dengan kata lain, distribusi kemampuan *posterior* yang diperbaharui pada proses kalibrasi digunakan sebagai distribusi *prior* pada proses kalibrasi berikutnya, sehingga perbedaan antara distribusi *prior* dan *posterior* menjadi minimal.

Data yang diamati pada estimasi kebolehjadian maksimum marjinal dianggap sebagai sampel acak dari suatu populasi. Prosedur estimasi kebolehjadian maksimum marjinal dikembangkan berdasarkan algoritma *Expectation-Maximization* (EM). Struktur dimensi dari butir individual didasarkan pada banyaknya kemampuan yang didefinisikan secara apriori dan memberikan kontribusi pada respons dengan menggunakan estimasi kebolehjadian maksimum marjinal yang menggunakan algoritma *Expectation-Maximization* (EM). Algoritma ini membandingkan peserta tes yang menjadi sampel acak dari populasi dan mengasumsikan tingkat kemampuan laten peserta berasal dari populasi yang berdistribusi normal baku, yaitu dengan nilai rata-rata 0 dan simpangan baku 1.

Prosedur EM merupakan prosedur yang berulang, banyaknya harapan (ekspektasi) peserta tes dari setiap tingkat kemampuan dihitung terlebih dahulu dengan banyaknya harapan dari orang yang menjawab benar setiap butir. Algoritma EM mempunyai tahapan ekspektasi dan tahap maksimalisasi. Estimasi parameter butir dilakukan dengan memaksimumkan kebolehjadian dari respons butir. Parameter butir digunakan untuk mengestimasi ulang frekuensi harapan, yang kemudian ditempatkan sekali lagi dalam persamaan kebolehjadian maksimum marjinal, dan proses ini berulang berkali-kali.

Beberapa kelebihan dan kekurangan metode estimasi kebolehjadian maksimum marjinal diuraikan berikut ini. Kelebihan metode estimasi kebolehjadian maksimum marjinal adalah: (1) dapat dipakai pada semua model IRT, termasuk model multidimensional; (2) efisien untuk tes berukuran panjang maupun pendek; (3) kebolehjadian maksimum marjinal mengestimasi kesalahan baku butir sebagai estimasi terbaik dari variansi penarikan sampel yang diharapkan; (4) estimasi dapat dilakukan terhadap skor sempurna, tidak dibutuhkan reduksi terhadap data yang mempunyai skor sempurna; (5) data kebolehjadian maksimum dapat digunakan untuk menguji hipotesis dan mengindikasi kecocokan data.

Adapun kekurangan metode estimasi kebolehjadian maksimum marjinal, adalah (1) algoritma yang efektif dalam mengestimasi sukar untuk diprogram; dan (2) distribusi kemampuan harus diasumsikan, sehingga membuat estimasi parameter tergantung pada ketepatan distribusi yang diasumsikan. Distribusi yang diasumsikan tidak perlu berdistribusi normal, tapi dapat diestimasi dari data. Berbagai tingkat distribusi kemampuan dapat dipertimbangkan, misalnya kurtosis miring dan tidak normal (Embretson & Reise, 2000, p.214).

Prosedur iterasi merupakan rangkaian berulang yang disebut langkah E (*expectation*) yang kemudian diikuti oleh langkah M (*maximization)* hingga mencapai konvergen. Jika sebanyak *n* butir tes diujikan pada N peserta dan respons butir dari peserta *i* terhadap butir *j* dinotasikan sebagai y_{ij} dan vektor respon butir amatan untuk *n* butir pada peserta *i* dapat dinotasikan sebagai $y_i = (y_{i1} \dots y_{ij} \dots y_{in})^T$, maka matriks data amatan dari N merupakan matriks $N \times n$ dan matriks $\mathbf{Y} = (y_1 \dots y_N)^T$. Misalkan Θ_i adalah variabel acak yang menyatakan variabel kemampuan laten pada peserta *i*, maka $\Theta_i = (\Theta_1 \dots \Theta_N)^T$ (McLachlan & Krishnan, 2008, pp.3-4). Asumsi ini digunakan juga pada pendekatan estimasi kebolehjadian maksimum marjinal dengan menggunakan *Expectation-Maximization*, dimana Θ_i adalah sampel acak dari suatu populasi.

Implikasi hasil temuan dalam penelitian simulasi Kim (2006) yang dirancang ideal memungkinkan untuk penelitian lebih lanjut dalam membandingkan kelima metode dengan menggunakan data riil. Kondisi riil dapat diperoleh dari data respons ter-

hadap perangkat tes yang memuat *common items*. Perangkat soal ujian nasional mata pelajaran matematika SMP tahun pelajaran 2009/2010 dikembangkan berdasarkan desain *anchor-test*. Desain ini merupakan desain yang dinilai tepat untuk kepentingan kalibrasi *fixed parameter* yang diaplikasikan pada data riil

Ada tiga model logistik dalam teori respos butir, yaitu model logistik satu, dua, dan tiga parameter. Model-model ini sesuai untuk data respons butir yang diskor dikotomus (Hambleton, Swaminathan, & Rogers, 1991, p.12). Perbedaan ketiga model ini adalah banyaknya parameter yang digunakan untuk menggambarkan karakteristik butir pada setiap model logistiknya atau parameter-parameter butir. Parameter-parameter butir tersebut adalah indeks kesukaran butir, indeks daya beda butir, dan tebakan semu. Ketiga unsur ini berhubungan sehingga menghasilkan fungsi atau lengkungan respons yang disebut juga kurva karakteristik butir.

Ujian nasional termasuk dalam tes kemampuan yang menggunakan format pilihan ganda, dimana model logistik tiga parameter cocok untuk digunakan. Peserta tes cenderung untuk memilih jawaban terbaik yang mereka anggap paling menarik jika mereka tidak dapat menemukan jawabannya. Sehingga faktor menebak dipertimbangkan dalam penelitian ini. Model logistik tiga parameter dinyatakan sebagai berikut.

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

(Hambleton, et al. ,1991; Embretson & Reise, 2000; Partchev, 2004). Pada model logistik tiga parameter, probabilitas peserta tes dalam menjawab benar suatu butir ditentukan oleh semua parameter butir, yaitu parameter kesukaran butir, parameter daya beda butir, dan *pseudo-guessing* (tebakan semu). Model logistik tiga parameter adalah model yang paling umum dari ketiga model. Dengan kurva berbentuk S dan asimptut yang lebih rendah, model ini sangat tepat ketika individu dengan kemampuan rendah terkadang dapat merespons dengan benar butir yang sulit (Hulin, Drasgow, & Parsons, 1983, p.29).

Penelitian ini bertujuan untuk (1) mengidentifikasi karakteristik butir-butir tes pada perangkat soal ujian nasional mata pelajaran Matematika tingkat SMP tahun pelajaran 2009/2010 yang dikalibrasi dengan metode kalibrasi *fixed parameter*, dan (2) mengetahui metode kalibrasi *fixed parameter* yang paling akurat dalam mengestimasi parameter di antara metode NWU-OEM, NWU-MEM, OWU-OEM, OWU-MEM, dan MWU-MEM.

Hasil penelitian ini dapat menjadi dasar pertimbangan bagi pengembang tes dalam menempatkan suatu butir bersama dalam tes yang akan dikembangkan. Hasil penelitian ini juga dapat menjadi rujukan untuk pengembangan bank soal.

Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif deskriptif. Penelitian dilakukan dengan menggunakan data respons ujian nasional mata pelajaran matematika tingkat SMP tahun pelajaran 2009/2010 dari provinsi DI Yogyakarta. Jumlah seluruh peserta Ujian Nasional tingkat SMP/ sederajat di Provinsi DI Yogyakarta tahun pelajaran 2009/2010 adalah 49.128. Jumlah respons yang menjadi sampel adalah 2500 respons siswa terhadap paket soal nomor 15 dan 2500 respons siswa terhadap paket soal nomor 48. Jumlah ini telah memenuhi syarat ukuran sampel minimal jika menggunakan model logistik tiga parameter (Hulin, Drasgow, & Parsons, 1983, pp.99-105).

Prosedur kalibrasi *fixed parameter* diawali dengan menetapkan parameter butir yang menjadi butir pengait atau *common items*. Parameter *common items* diperoleh dari estimasi yang dilakukan terhadap data respons perangkat tes paket 15. Parameter *common items* kemudian ditempatkan sebagai *fixed parameter* pada saat proses kalibrasi terhadap data respons perangkat tes paket 48. Berikut adalah prosedur kerja setiap metode.

(1) Pada metode NWU-OEM, kemampuan awal tidak diperbaharui. Kebolehjadian posterior untuk setiap peserta tes hanya dipengaruhi oleh *common items*, tidak oleh butir-butir baru sama sekali. Metode ini menggunakan hanya satu langkah E, yang hanya melibatkan data respons dari *common items*. Metode ini juga hanya terdiri dari satu langkah M dan melibatkan data respons dari butir-butir yang baru.
(2) Pada metode NWU-MEM, siklus EM yang pertama dilakukan dengan cara yang sama seperti metode NWU-OEM. Pada langkah kedua E dari metode ini, respons butir yang digunakan berasal dari *common items* dan butir-butir baru. Respons butir untuk butir-butir yang baru digunakan sebagai nilai posterior kebolehjadian. Banyaknya siklus EM yang ditetapkan adalah 2000.
(3) Pada metode OWU-OEM, siklus tunggal EM diaplikasikan satu kali pada butir-butir yang baru, ketika kemampuan awal telah diperbaharui sebanyak satu kali. Nilai posterior dihitung setelah satu kali siklus EM. Seperti halnya metode NWU-OEM, pada metode ini hanya *common items* yang digunakan untuk menghitung kebolehjadian bersyarat untuk setiap peserta tes.
(4) Metode OWU-MEM mengestimasi parameter pada butir-butir baru melalui multi siklus EM dengan memperbaharui distribusi kemampuan sebanyak satu kali. Siklus EM yang pertama dijalankan dengan cara yang sama dengan metode NWU-OEM, dan nilai prior yang diperbaharu, digunakan pada siklus EM yang kedua dan diusahakan tidak berubah selama siklus EM berikutnya. Kemampuan awal diperbaharui kembali sebanyak satu kali lagi setelah siklus EM terakhir. Sama dengan metode NWU-MEM, *common items* dan butir-butir baru digunakan untuk mengestimasi distribusi kemampuan. Banyaknya siklus EM yang ditetapkan adalah 2000.
(5) Metode MWU-MEM memperbaharui distribusi kemampuan awal dan menemukan estimasi kebolehjadian maksimum pada parameter butir-butir yang baru secara berulang. Siklus EM yang pertama dilakukan dengan cara yang sama seperti pada metode NWU-OEM. Langkah berikutnya adalah memperbaharui nilai awal yang ditetapkan. Pada siklus EM yang kedua kali, parameter butir dan nilai kemampuan diestimasi secara bersama-sama. Setelah siklus EM yang terakhir, diproleh nilai posterior laten. Metode MWU-MEM menggunakan semua informasi dari respons butir-butir yang baru untuk memperoleh distribusi kemampuan laten dan untuk mengestimasi parameter. Banyaknya siklus EM yang ditetapkan adalah 2000.

Data respons ujian nasional mata pelajaran Matematika SMP dipilih sebagai sumber data karena dalam penelitian ini diperlukan perangkat tes yang memuat sejumlah *common items*. Data respons Ujian Nasional merupakan data dokumentasi. Data ini diperoleh dari pusat data Pusat Penilaian Pendidikan Kemendikbud RI. Perangkat tes terdiri dari dua paket yang berasal dari kisi-kisi tes yang sama, yaitu paket nomor 15 dan 48. Setiap paket berisi 40 butir soal bentuk pilihan ganda dengan empat pilihan jawaban. Siswa dan responsnya terhadap paket soal nomor 15 menjadi kelompok *base* dan siswa dan respons terhadap paket soal nomor 48 menjadi kelompok *target*. Estimasi terhadap parameter butir paket soal nomor 15 dilakukan untuk memperoleh parameter-parameter *common items*. Parameter-parameter *common items* selanjutnya akan menjadi *fixed-parameter* pada proses estimasi selanjutnya terhadap paket soal nomor 48.

Tes pada desain ini digambarkan pada Gambar 1. Tampak dua perangkat tes yang digabungkan dengan 6 butir inti. Jumlah soal seluruhnya setiap perangkat adalah 40 butir. Pada masing-masing perangkat terdiri dari 6 butir inti dan 34 butir yang bukan butir inti. Enam butir inti ini adalah soal yang menghubungkan perangkat tes 15 dengan perangkat tes 48. Soal inti yang menghubungkan setiap perangkat berfungsi sebagai *common items*.

P 15		P 48
34	6	34

Keterangan: P : Perangkat

Gambar 1. Desain Tes yang Memuat *Common Items*

Akurasi setiap metode dievaluasi dengan memperhatikan fungsi informasi tes (TIF) dan kesalahan baku pengukuran (SEM) pada setiap metode. Fungsi informasi butir merupakan fungsi yang memberikan informasi tentang sumbangan butir dalam mengungkap kemampuan laten yang diukur oleh suatu tes. Melalui fungsi ini dapat diketahui butir mana yang cocok dengan model, sehingga dapat membantu dalam penyeleksian butir soal.

Fungsi informasi tes (TIF) merupakan jumlah dari fungsi informasi butir-butir dalam tes. Fungsi informasi tes secara matematis dinyatakan sebagai berikut.

$$I(\hat{\theta}) = \sum_{i=1}^{n} I_i(\theta)$$

(Hambleton, et al. ,1991, 94; Embretson & Reise, 2000, p.184; Partchev, 2004, p.16). Nilai Fungsi informasi tes akan besar jika butir-butir penyusun tes memiliki nilai fungsi informasi butir yang besar pula.

Kesalahan pengukuran digambarkan dalam unit yang sama sebagai pengukuran itu sendiri, hal ini dapat dibandingkan dengan estimasi kemampuan, atau menggunakannya untuk membangun interval kepercayaan di seputar estimasi. Varians dari estimasi kemampuan $\hat{\theta}$ dapat diestimasi sebagai nilai resiprokal dari fungsi informasi tes terhadap $\hat{\theta}$, yang dinyatakan dalam persamaan berikut.

$$Var\,(\hat{\theta}) = \frac{1}{I(\hat{\theta})}$$

Kesalahan baku pengukuran berkaitan dengan fungsi informasi, yaitu berbanding terbalik secara kuadratik (Hambleton & Jones, 1993). Dalam teori respons butir, setiap skor yang diestimasi memiliki kesalahan baku pengukuran. Kesalahan baku pengukuran (*SEM*) juga merupakan akar kuadrat dari varians, maka secara matematis hubungan ini dapat dituliskan berikut.

$$SEM\,(\theta) = \sqrt{\frac{1}{I(\hat{\theta})}}$$

Persamaan di atas menunjukan bahwa semakin besar fungsi informasi tes dan semakin kecil kesalahan baku pengukuran suatu metode memberi indikasi metode tersebut semakin akurat dalam mengestimasi parameter butir.

Software yang digunakan sebagai alat bantu dalam estimasi dan analisis hasil penelitian ini adalah *BILOG-MG, PARSCALE*, dan *Excell for Windows.* Menurut Crocker & Algina (1986, p.354), Hambleton (1991, pp.42-50), dan Yen & Fitzpatrick (2006, pp.131-132), program komputer untuk estimasi kebolehjadian maksimum adalah program *Bilog. Bilog-MG* fit untuk model logistik satu, dua, dan tiga parameter dengan menggunakan prosedur Marginal Maximum Likelihood. Program *Bilog-MG* dapat mengestimasi butir-butir bentuk pilihan ganda, dan dapat digunakan untuk skala besar dalam mengestimasi kemampuan latent. Selain Bilog-MG digunakan juga program *Parscale. Parscale* mengestimasi parameter dari model respons berdasarkan Marginal Maximum Likelihood (du Toit, 2003, p.611). *Parscale* digunakan untuk mengkalibrasi *fixed parameter* berdasarkan distribusi kebolehjadian maksimum atau distribusi posterior dengan algoritma EM. Program *Bilog-MG* dan *Parscale* menggunakan *multiple* siklus EM, namun program *Bilog-MG* tidak dapat men-*update* distribusi kemampuan prior seperti pada program *Parscale* (Taehoon & Petersen, 2009 dan Kim, 2006). Dengan dasar itulah, maka pada penelitian ini untuk mengestimasi parameter dengan metode NWU-OEM dan metode NWU-MEM menggunakan program *Bilog-MG* dan program *Parscale* digunakan untuk mengestimasi parameter dengan metode OWU-OEM, OWU-MEM, dan MWU-MEM.

Hasil Penelitian dan Pembahasan

Hasil Penelitian

Kedua paket terhubung oleh 6 buah *common items*. Hal ini dapat dilihat pada Tabel 1. Tabel 1 menunjukkan terdapat 6 buah butir soal yang menjadi butir pengait (*common items)* kedua paket soal. Butir-butir tersebut ditempatkan pada nomor urut yang sama, kecuali butir soal nomor 29 pada Paket 15 ditempatkan sebagai butir soal nomor 22 pada Paket 48.

Tabel 1. Sebaran *Common Items*

Perangkat Soal	Nomor Soal					
Paket 15	8	25	26	27	29	31
Paket 48	8	25	26	27	22	31

Analisis terhadap tingkat kesukaran butir pada paket 15 dan 48 menunjukkan terdapat 16 butir yang mudah, 24 butir yang tingkat kesukarannya sedang, dan tidak ada butir yang tergolong sukar pada paket 15. Hasil analisis butir pada paket 48 menunjukkan terdapat 19 butir yang tergolong mudah, 20 butir yang tingkat kesukarannya sedang, dan 1 butir yang tergolong sukar, yaitu butir nomor 21.

Analisis terhadap koefisien korelasi biserial mengacu pada Mardapi (2012, 129), yaitu butir soal yang diterima adalah butir soal yang memiliki koefisien korelasi biserial ≥ 0,30. Hasil analisis terhadap paket 15 diperoleh 2 butir soal memiliki koefisien korelasi biserial kurang dari 0,30, yaitu butir soal nomor 3 dengan koefisien korelasi sebesar 0,23 dan butir soal nomor 28 dengan koefisien korelasi sebesar 0,21; sedangkan pada paket 48 diperoleh hanya ada satu butir soal memiliki koefisien korelasi biserial kurang dari 0,30, yaitu butir soal nomor 21 dengan koefisien korelasi sebesar 0,02.

Hasil analisis kecocokan model menunjukkan bahwa kedua paket soal fit dengan model logistik tiga parameter, karena terdapat lebih dari 80% butir pada setiap paket soal yang cocok dengan model. Butir yang memiliki indeks daya beda yang paling tinggi pada paket 15 terdapat pada butir soal nomor 17 yaitu 1,60 sedangkan pada paket 48 butir 34 memiliki indeks daya beda butir paling tinggi yaitu 1,76. Indeks daya beda terkecil pada paket 15 terdapat pada butir soal nomor 33, yaitu 0,48 dan butir soal nomor 28 pada paket 48, yaitu 0,61. Semua butir pada paket 15 dan 48 berada pada interval [-3, 3], butir soal yang mempunyai indeks kesukaran terkecil pada paket 15 adalah butir soal nomor 33, yaitu -2,27 dan butir soal nomor 28 pada paket 48, yaitu -2,14. Indeks kesukaran butir paling besar pada paket 15 terdapat pada butir soal nomor 38, yaitu 1,19 dan pada paket 48 terdapat pada butir soal nomor 17, yaitu 1,06.

Terdapat 12 butir pada paket 15 dan 19 butir pada paket 48 yang memiliki indeks parameter *pseudoguessing* yang lebih dari 0,25.

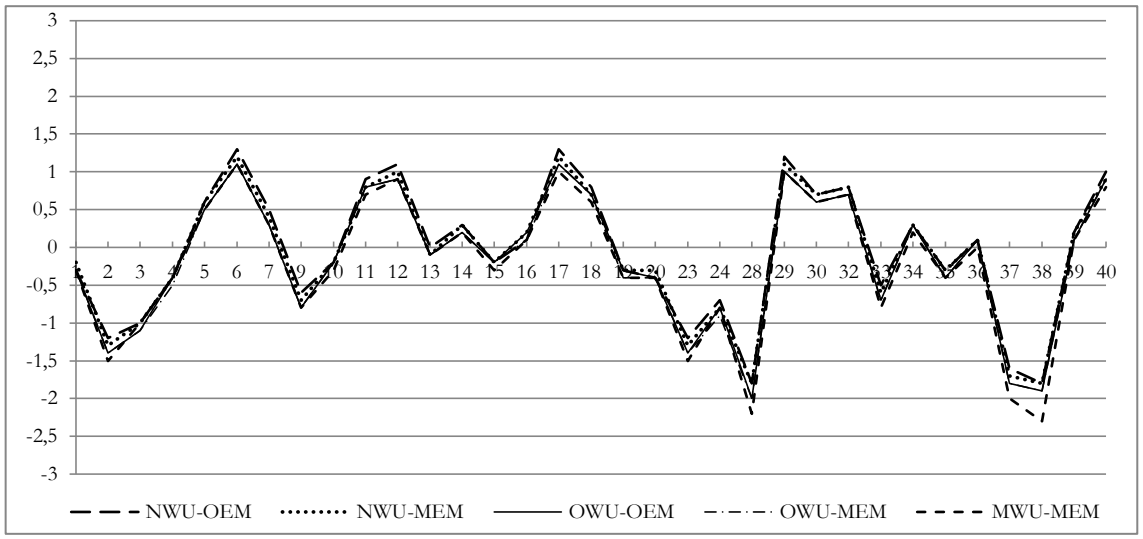
Analisis faktor dilakukan untuk mengetahui apakah asumsi unidimensi terpenuhi. Untuk mengetahui muatan faktor pada kedua perangkat tes dilakukan dengan menggunakan program *SPSS 16.0 for Windows*. Nilai KMO untuk paket 15 adalah 0,962 dan untuk paket 48 adalah 0,905. Nilai ini mendekati 1, menunjukkan bahwa pola korelasi relatif kompak dan juga faktor analisis menghasilkan faktor-faktor yang berbeda. Hasil analisis faktor tentang kecukupan sampel menunjukkan nilai Khi-kuadrat pada uji *Bartlett* paket 15 adalah 19380 dengan derajat bebas 780 dan nilai $p < 0{,}001$; sedangkan nilai Khi-kuadrat untuk paket 48 adalah 12350 dengan derajat bebas 780 dan nilai $p < 0{,}001$. Hasil ini menunjukkan bahwa ukuran sampel sebesar 2.500 yang digunakan pada penelitian ini telah mencukupi.

Berdasarkan nilai *eigen* pada paket 15 terdapat 6 faktor yang memiliki nilai lebih dari 1 dan pada paket 48 terdapat 10 faktor. Dengan mengambil kriteria nilai *eigen* pertama minimal 20% dari total nilai *eigen* faktor dominan (Reckase, 1979), maka dapat dikatakan bahwa butir-butir soal pada kedua paket soal telah memenuhi asumsi unidimensi.

Hasil estimasi terhadap butir-butir paket 15 menghasilkan parameter *common*

items yang selanjutnya ditempatkan sebagai *fixed-parameter* untuk mengestimasi parameter butir pada Paket 48. Hasil estimasi terhadap *common items* pada paket 15, menunjukkan bahwa parameter kesukaran butir berada pada interval [-1,11, 0,74], parameter daya beda butir berada pada interval [0,63, 1,34], dan parameter *pseudo-guessing* memiliki nilai c kurang dari 0,25, kecuali butir soal nomor 26, yaitu 0,46 dan 0,45.

Tampak pada Gambar 2 nilai theta atau nilai kemampuan pada posisi fungsi informasi butir menjadi maksimal yang ditunjukkan oleh kelima metode kalibrasi *fixed-parameter* sangat kecil perbedaannya, kelima grafik fungsi hampir berimpit. Metode NWU-OEM mempunyai rerata fungsi informasi butir yang paling kecil dan metode OWU-OEM mempunyai rerata fungsi informasi butir terbesar diantara kelima metode ini.



Gambar 2. Grafik Nilai *Theta* dimana Fungsi Informasi Butir menjadi Maksimal

Tabel 2. Rerata Parameter Butir dan Parameter *Non-common Items, TIF*, dan *SEM* pada setiap Metode

Metode	Rerata Parameter Butir			TIF	SEM
	a	*b*	*c*		
NWU-OEM	1,07	-0,20	0,23	4,90	0,4532
NWU-MEM	1,14	-0,23	0,23	5,16	0,4404
OWU-OEM	1,14	-0,29	0,22	5,30	0,4343
OWU-MEM	1,12	-0,30	0,22	5,22	0,4376
MWU-MEM	1,13	-0,35	0,23	5,12	0,4418
Metode	**Rerata Parameter *Non-common items***			**TIF**	**SEM**
	a	***b***	***c***		
NWU-OEM	1,09	-0,20	0,24	4,15	0,4908
NWU-MEM	1,17	-0,23	0,24	4,44	0,4747
OWU-OEM	1,17	-0,28	0,23	4,55	0,4686
OWU-MEM	1,15	-0,29	0,23	4,47	0,4728
MWU-MEM	1,16	-0,35	0,24	4,37	0,4782

Keterangan:
TIF : *Test Information Function* (Fungsi Informasi Tes)
SEM : *Standard Error Measurement* (Kesalahan Baku Pengukuran)

Tabel 2 menunjukkan rerata indeks daya beda butir berada pada interval [1,07, 1,14] dengan rentangan rerata yang sangat kecil yaitu 0,07 dan rerata indeks kesukaran butir berada pada interval [-0,35, -0,20] dengan rentangan 0,15. Rerata indeks *pseudo guessing* pada setiap metode kurang dari 0,25. Rerata indeks daya beda butir parameter *non-common-items* berada pada interval [1,08, 1,17] dengan rentangan 0,09. Rerata indeks kesukaran butir berada pada interval [-0,35, -0,20] dengan rentangan 0,15. Rerata parameter *pseudoguessing* pada setiap metode memiliki indeks kurang dari 0,25.

Fungsi informasi tes tertinggi pada parameter butir dan parameter *non-common*

items terdapat pada metode OWU-OEM dan yang paling rendah adalah metode NWU-OEM. Seiring dengan fungsi informasi tes, indeks kesalahan baku pengukuran pada parameter butir dan parameter *non-common items* paling rendah terdapat pada metode OWU-OEM dan yang paling tinggi adalah metode NWU-OEM. Berdasarkan fungsi informasi tes dan kesalahan baku pengukuran maka metode OWU-OEM merupakan metode yang paling akurat dalam mengestimasi parameter baik parameter butir maupun parameter *non-common items*. Jika diurutkan berdasarkan fungsi informasi tes dan kesalahan baku pengukuran, maka urutan tingkat akurasi metode dalam mengestimasi parameter adalah metode OWU-OEM, OWU-MEM, NWU-MEM, MWU-MEM, dan metode NWU-OEM.

Pembahasan

Common items mempunyai peranan penting dalam proses kalibrasi butir. Sifat dan karakteristik *common items* menggambarkan miniatur suatu perangkat tes yang di dalamnya memuat sejumlah *common items*. Mencermati butir-butir yang menjadi *common items* pada perangkat tes ujian nasional mata pelajaran Matematika tahun pelajaran 2009/2010 tampak bahwa jumlah *common items* adalah 6 butir yaitu butir soal nomor 8, 25, 26, 27, 29 (22), dan 31. Jumlah ini masih sangat sedikit jika mengacu kepada Kolen dan Brennan (1995, p.248), yaitu minimal 20% dari panjang tes. Jika mengacu pada hal tersebut, maka seharusnya jumlah *common items* paling sedikit 8 butir dari 40 butir yang terdapat pada perangkat tes ujian nasional mata pelajaran matematika.

Penempatan butir-butir soal yang menjadi *common items* pada nomor urut yang sama, baik pada paket soal 15 dan paket soal 48 telah memenuhi kriteria yang disarankan oleh para ahli. Semua butir soal ditempatkan pada nomor soal yang sama pada kedua paket soal, kecuali butir soal nomor 29 pada paket soal nomor 15 ditempatkan sebagai butir soal nomor 22 pada paket soal nomor 48.

Sifat dan karakteristik *common items* menggambarkan miniatur suatu perangkat tes juga dapat dilihat pada penyebaran materi dan tingkat kelas yang diwakili oleh *common items*. Hal tersebut disajikan pada Tabel 3.

Tabel 3. Sebaran Materi pada *Common Items*

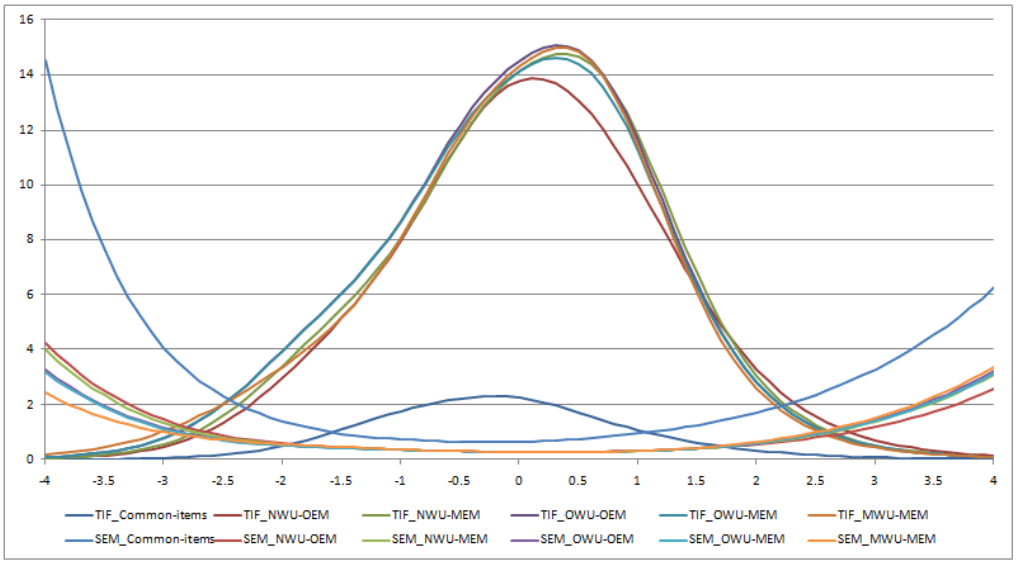
Nomor Soal	Indikator	Materi	Kelas
8	Mengalikan bentuk Aljabar	Bentuk Aljabar	VII dan VIII
25	Menghitung besar sudut yang terbentuk jika dua garis sejajar berpotongan dengan garis lain	Garis-Garis Sejajar	VIII
26	Menghitung besar sudut yang melibatkan sudut dalam dan sudut luar segitiga	Sudut	VII
27	Menghitung besar sudut pusat atau sudut keliling pada lingkaran	Lingkaran	VIII
29 (22)	Menyelesaikan masalah dengan menggunakan konsep kesebangunan	Kesebangunan	IX
31	Menentukan unsur-unsur pada kubus dan balok	Kubus dan Balok	VIII

Tampak butir-butir soal yang menjadi *common items* hanya mencakup materi Aljabar dan Geometri pada semua tingkatan kelas. Tidak semua aspek materi terwakili oleh butir-butir soal yang menjadi *common items*. Jumlah *common items* yang sedikit menyebabkan tidak semua materi dapat terwakili.

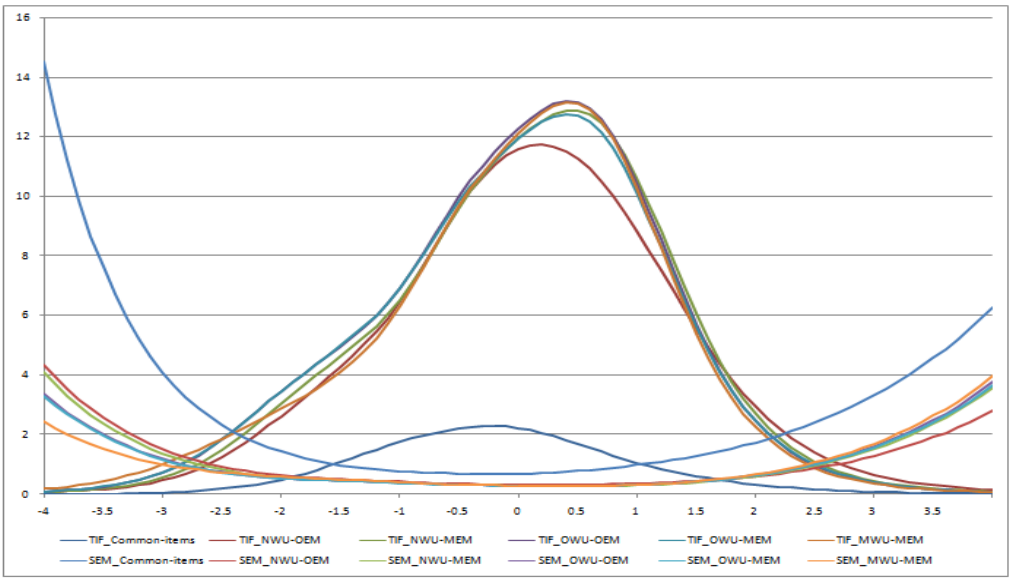
Analisis terhadap kesesuaian butir soal dengan Standar kompetensi lulusan (SKL), materi, dan indikator (Depdiknas, 2006) menunjukkan beberapa indikator yang telah ditetapkan tidak dibuatkan butir soalnya. Indikator tersebut adalah (1) mengurutkan pecahan jika diberikan beberapa

jenis pecahan, (2) menentukan penyelesaian persamaan linear satu variabel dalam bentuk pecahan, (3) menghitung luas gabungan dua bangun datar, dan (4) menghitung luas juring lingkaran dari unsur yang diketahui. Terdapat juga dua butir soal yang tidak sesuai dengan indikator yang diberikan, yaitu butir soal nomor 11 (15) dan nomor 24 (36). Hal ini mungkin disebabkan SKL dan indikator yang diberikan kepada sekolah dan siswa berlaku untuk seluruh paket soal di seluruh daerah di Indonesia, jadi memungkinkan terjadi tidak semua indikator dibuatkan soalnya pada suatu daerah tapi ada pada paket soal di daerah yang lain.

Fungsi informasi tes dan kesalahan baku pengukuran pada semua butir serta fungsi informasi tes dan kesalahan baku pengukuran pada *non-common items* dirangkum pada Gambar 3 dan 4. Kedua gambar memperkuat hasil analisis yang menunjukkan bahwa *common items* tidak cukup mewakili kemampuan yang akan diukur. Hal ini disebabkan oleh jumlah butir yang tidak mencukupi dan sebaran materi yang hanya mencakup dua aspek materi, yaitu Aljabar dan Geometri.



Gambar 3. Grafik TIF dan SEM Semua Butir



Gambar 4. Grafik TIF dan SEM *Non Common Items*

Hasil analisis terhadap nilai theta di posisi fungsi informasi butir menjadi maksimal menunjukkan grafik fungsi kelima metode kalibrasi *fixed parameter* hampir berimpit. Metode OWU-OEM mempunyai rerata fungsi informasi butir paling besar di antara kelima metode kalibrasi *fixed parameter*, sehingga dapat dikatakan bahwa butir-butir yang dikalibrasi dengan metode OWU-OEM memberikan lebih banyak informasi hasil pengukuran dibandingkan metode lainnya. Berdasarkan fungsi informasi tes dan kesalahan baku pengukuran terkecil, metode OWU-OEM juga merupakan metode yang memiliki performa paling akurat dalam mengestimasi parameter.

Hasil penelitian ini menunjukkan hasil yang berbeda dengan hasil penelitian yang telah dilakukan Kim (2006), bahwa metode MWU-MEM menunjukkan performa yang paling akurat pada tiga distribusi kemampuan normal yang ditetapkan, yaitu N(0,1), N(0,5;1,2^2), dan N(1; 1,4^2). Hasil analisis terhadap akurasi terhadap kelima metode dalam penelitian ini, menunjukkan metode OWU-OEM merupakan metode yang paling akurat dalam mengestimasi parameter butir maupun parameter *non-common items*. Hal ini dapat dijelaskan karena data yang digunakan dalam penelitan ini adalah data riil yang memiliki karakteristik data yang secara teoritis berbeda dengan data yang dikembangkan dalam penelitian simulasi Kim.

Simpulan

Berdasarkan hasil penelitian dan pembahasan, diperoleh kesimpulan sebagai berikut: (1) statistik parameter butir-butir tes pada perangkat ujian nasional mata pelajaran matematika tingkat SMP tahun pelajaran 2009/2010 menunjukkan rerata indeks daya beda butir berada pada interval [1,07, 1,14], rerata indeks kesukaran butir [-0,35, -0,20], dan rerata *pseudo guessing* < 0,25. Nilai theta pada posisi fungsi informasi butir menjadi maksimal menunjukkan grafik fungsi kelima metode kalibrasi *fixed parameter* hampir berimpit. (2) metode OWU-OEM merupakan metode yang paling akurat dalam mengestimasi parameter butir pada perangkat tes ujian nasional mata pelajaran Matematika tahun pelajaran 2009/2010.

Implikasi hasil penelitian ini berkenaan dengan pengembangan bank soal, yaitu pada proses pengembangan bank soal dibutuhkan ketepatan pemilihan metode kalibrasi agar diperoleh estimasi parameter butir yang akurat. Pelaksanaan ujian nasional tahun pelajaran 2012/2013 lalu membutuhkan paling sedikit 20 paket soal atau setara dengan 800 butir soal untuk satu daerah, sehingga kebutuhan akan keberadaan sejumlah besar butir soal yang telah dikalibrasi menjadi sangat penting. Hasil penelitian ini memberikan informasi bagaimana memperoleh butir soal terkalibrasi dengan menggunakan varians metode kalibrasi *fixed parameter*. Hasil penelitian ini juga memberikan beberapa alternatif metode dalam proses kalibrasi butir soal dan menempatkan sejumlah *common items* yang dapat meningkatkan akurasi estimasi parameter butir yang dikembangkan untuk kepentingan pelaksanaan kegiatan ujian di daerah.

Penelitian ini hanya dilakukan pada dua perangkat tes ujian nasional mata pelajaran matematika, sehingga disarankan perlu ada penelitian lebih lanjut terhadap perangkat ujian nasional yang memiliki lebih dari dua paket soal dan juga pada mata pelajaran lain selain matematika.

Daftar Pustaka

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ban, J-C., Hanson, B.A., Tianyou Wang, et al. (2001) A comparative study of online pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, 38, 191-212.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New

York: Holt, Rinehart and Winston Inc.

Depdiknas. (2000). *Penilaian dan pengujian untuk guru SLTP*. Jakarta: Direktorat Jenderal Pendidikan Dasar dan Menengah, Direktorat Sekolah Lanjutan Tingkat Pertama, Depdiknas.

Depdiknas. (2005). *Peraturan Pemerintah RI Nomor 19, Tahun 2005, tentang Standar Nasional Pendidikan.*

Depdiknas. (2006). *Peraturan Menteri Pendidikan Nasional RI Nomor 23, Tahun 2006, tentang Standar Kompetensi Lulusan untuk Satuan Pendidikan Dasar dan Menengah.*

du Toit, M. (Ed.) (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publisher.

Green, D.R., Yen, W.M., & Burket,G.R. (1989). Experiences in the application of item response theory in test construction. *Journal of Educational Measurement*, 2, 297-312.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

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

Hambleton, R. K & Jones, R. W. (1993). *An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development*. Diambil pada tanggal 5 Juli 2010, dari www.ncme.org/pubs/items/10. *pdf*.

Hulin, C.L., Drasgow, F. & Parsons, C.K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, 43, 355-381.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer.

Lee, W-C & Ban, J-C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education,* 23, 23-48.

Li,-Y. H., Griffith, W. D., & Tam, H.P. (1997, June). *Equating multiple tests via an IRT linking design: Utililizing a single set of common items with fixed common item parameters during the calibration process.* Paper presented at the annual meeting of the psychometric society, Knoxville, TN.

Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Yogyakarta: Nuha Medika.

McLachlan, G.J. & Krishnan, T. (2008). *The EM algorithm and extensions (2nd ed.)*. New York: John Wiley & Sons.

Partchev, I. (2004). *A visual guide to item response theory*. Diambil pada tanggal 12 April 2009.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.

Taehoon Kang & Petersen, N. (2009). Linking item parameters to a base scale. *ACT Research Report Series*, 2009-2. Diambil tanggal 20 September 2010, dari http://www. act.org/research/researchers/reports/pdf/ACT_RR2009-2.pdf.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Yen, W. M & Fitzpatrick, A. R. (2006). Item response theory dalam R.L. Brennan (Ed.), *Educational measurement. 4th ed.* (pp.111-153). Westport, CT: American Council on Education and Praeger Publishers.