HUBUNGAN MODEL PENSKORAN TERHADAP ESTIMASI SKOR SESUNGGUHNYA BERDASARKAN TEORI RESPONS BUTIR

Musmuliadi
LPMP NTB
musmuliadi1@yahoo.co.id

Abstrak

Penelitian ini bertujuan untuk mendeskripsikan: 1) karakteristik tes UN mata pelajaran matematika tingkat SLTP tahun 2007/2008, 2) karakteristik distribusi skor sesungguhnya hasil estimasi beberapa model penskoran, 3) hubungan antara skor kemampuan (θ) dan skor tampak dengan skor sesungguhnya, dan 4) implikasi penerapan model penskoran terhadap estimasi skor sesungguhnya. Data penelitian ini berupa respons siswa SMP/MTs terhadap tes Ujian Nasional (UN) mata pelajaran matematika tahun 2007/2008 di Propinsi Nusa Tenggara Barat. Analisis dilakukan dengan pendekatan kuantitatif. Hasil analisis menunjukkan bahwa tes UN mata pelajaran matematika tahun 2007/2008 tingkat SMP/MTs pada kategori sulit, memiliki rerata daya pembeda baik, tetapi rerata indeks *pseudo-guessing* kurang baik. Rerata skor sesungguhnya yang paling tinggi diperoleh pada model penskoran jumlah benar sesungguhnya, sedangkan rerata paling kecil terjadi pada model penskoran koreksi terhadap tebakan. Hubungan antara skor kemampuan (θ) dengan skor sesungguhnya menunjukkan korelasi positif dengan nilai koefisien korelasi sangat tinggi. Rerata hasil estimasi skor sesungguhnya dari ketiga model penskoran menunjukkan perbedaan yang signifikan.

Kata kunci: *skor tampak, skor kemampuan, skor sesungguhnya, model penskoran*

THE IMPLICATION OF FORMULA SCORING ON THE TRUE SCORE ESTIMATION BASED ON ITEM RESPONSE THEORY

Musmuliadi
LPMP NTB
musmuliadi1@yahoo.co.id

Abstract

This study is aimed to investigate: 1) the characteristics of Mathematics National Examination Test at junior high schools in 2007/2008 academic year, 2) the characteristics of the true score based on some formula scorings, 3) the relationship between observed score and ability score (θ) with true score, and 4) the implication of applying formula scoring to estimate true score. This research data were in the form of Mathematics National Examination responses at junior high schools in 2007/2008 academic year in Nusa Tenggara Barat province. The result of analysis indicates that Mathematics National Examination Test at junior high schools in 2007/2008 academic year had a high average of item difficulty, a good average of discrimination index, but bad average of pseudo-guessing index. The comparison among true scores shows that the highest average is found when using number of right true score formula and the smallest one obtained by correction for guessing formula, while the most disseminating true score distribution is found when using optimal weighting formula. The ability score (θ) and the true score correspond to a very high positive correlation coefficient. The true score average caused by all of the formulas were significantly different.

Key word: *observed score, ability score, true score, formula scoring*

Pendahuluan

Pengukuran dalam bidang pendidikan menyaratkan kuantifikasi suatu atribut yang diukur berdasarkan aturan tertentu. Pengukuran, dalam hal ini pengukuran terhadap hasil belajar, dilakukan pada setiap jenjang proses pendidikan dan merupakan proses yang esensial. Pengukuran ini dapat dilakukan pada awal program pembelajaran, selama program pembelajaran, dan pada akhir program pembelajaran. Alat ukur yang umum digunakan adalah tes di mana tes tersebut dapat berasal dari buatan guru atau menggunakan tes standar.

Bentuk tes yang paling umum digunakan adalah tes pilihan ganda karena tes berbentuk pilihan ganda dapat diskor dengan mudah dan cepat, objektif, lebih mudah dianalisis, dapat mencakup materi yang luas dalam suatu tes, dapat mengukur kemampuan yang bermacam–macam dari yang paling sederhana sampai yang paling kompleks (Nitko & Brookhart, 2007: 151-152; Gronlund & Linn, 1990: 174-177). Tetapi, tes pilihan ganda memberikan peluang lebih besar bagi peserta ujian untuk melakukan kecurangan (*cheating*) dalam ujian (Sotaridona, van der Linden, & Meijer, 2006: 412; van der Linden & Sotaridona, 2004: 361). Tes pilihan ganda juga sangat peka terhadap tebakan dan tidak sensitif terhadap perbedaan tingkat pengetahuan siswa (Simon, Budescu, & Nevo, 1997: 65).

Tebakan akan menghasilkan skor tes yang menyebabkan menurunnya reliabilitas dan validitas (Chopin 1988: 384). Selain itu, tebakan yang menguntungkan akan mendongkrak skor siswa sehingga menyebabkan terjadinya estimasi yang terlalu tinggi terhadap kemampuan siswa tersebut. Oleh karena itu, Sax (1980: 100) menyarankan agar menambah jumlah pilihan pada setiap butir soal dan menyesuaikan batas waktu ujian. Sementara Wang (1995: 4) mengusulkan supaya menetapkan nilai kritik kelulusan pada suatu tes (*passing score*) berdasarkan tebakan meskipun sangat kecil kemungkinan seorang siswa memperoleh nilai tinggi dengan menebak. Chopin (1988: 386) memberikan salah satu alternatif untuk mengurangi efek tebakan pada tes pilihan ganda dengan menerapkan model penskoran koreksi terhadap tebakan (*correction for guessing*).

Skor yang diperoleh seorang siswa sering dinyatakan sebagai estimasi kemampuan siswa tersebut terhadap materi yang diukur oleh suatu tes. Menurut Rowley & Traub (Crocker & Algina, 1986: 400), penentuan skor komposit suatu hasil tes pilihan ganda didasarkan pada tiga kondisi. Pertama, siswa mengetahui pilihan jawaban yang benar dan memilih pilihan tersebut. Kedua, siswa mengabaikan soal yang tidak bisa dijawabnya. Ketiga, siswa menebak salah satu dari *k* pilihan jawaban secara random.

Estimasi kemampuan siswa dapat dilakukan melalui pendekatan teori tes klasik (*Classical Test Theory,* selanjutnya ditulis CTT) maupun teori respons butir (*Item Response Theory*, selanjutnya ditulis IRT). Pendekatan teori tes klasik lebih umum digunakan dalam praktik karena kesederhanaan dalam perhitungan. Berdasarkan pendekatan ini kemampuan seorang siswa terhadap materi yang diukur oleh suatu tes pilihan ganda diestimasi dengan menggunakan jumlah butir yang dijawab benar dan sering dinyatakan sebagai skor tampak (X). Estimasi kemampuan berdasarkan pendekatan ini tidak memperhatikan pola respons siswa yang menjawab. Hasil estimasi kurang sensitif terhadap karakteristik butir. Karakteristik butir seperti tingkat kesulitan butir, daya pembeda butir, dan efek tebakan tidak dipertimbangkan dalam mengestimasi kemampuan siswa. Selain itu, bobot tiap butir soal dianggap sama.

Menurut Garcí–Pérez & Frary (1989: 403), skor jumlah benar pada tes pilihan ganda tidak beralasan diklaim sebagai estimasi kemampuan yang dimiliki oleh siswa. Metode ini hanya menunjukkan informasi tentang rangking siswa. Berbeda dengan teori tes klasik, estimasi kemampuan siswa pada IRT ditentukan berdasarkan pola responsnya, tidak ditentukan berdasarkan jumlah butir yang dijawab dengan benar. Jadi, pola respons siswa yang bervariasi menunjukkan variasi kemampuan siswa (Djemari Mardapi, 1999: 9-10; Thissen & Orlando, 2001: 114-117). Estimasi kemampuan siswa berdasarkan IRT sering dinyatakan sebagai θ.

Skor kemampuan atau θ yang dihasilkan oleh IRT diukur pada skala yang kurang lazim karena skor tersebut dapat bernilai negatif sehingga mempersulit sebagian orang untuk menginterpretasi. Agar lebih mudah diinterpretasi, skor kemampuan tersebut kemudian dinyatakan dalam bentuk skor sesungguhnya (*true score*). Konsep skor kemampuan dan skor

sesungguhnya pada dasarnya sama kecuali pada skala pengukuran kedua ukuran tersebut. Skor kemampuan didefinisikan pada interval $(-\infty, +\infty)$ sedangkan skor sesungguhnya didefinisikan pada interval $[0, n]$. Perbedaan utama kedua skala tersebut adalah bahwa skala kemampuan tidak tergantung pada jumlah butir soal dalam tes sementara skala skor sesungguhnya tergantung pada jumlah butir soal dalam tes (n).

Skor sesungguhnya merupakan besaran yang tidak bisa diukur secara langsung melainkan melalui estimasi. Estimasi skor sesungguhnya dapat dilakukan berdasarkan CTT maupun IRT. Estimasi skor sesungguhnya berdasarkan CTT didasarkan pada skor tampak siswa, rata-rata skor tampak peserta ujian seluruhnya, dan indeks reliabilitas tes yang digunakan pada ujian tersebut (Mehrens & Lehmann, 1973: 107). Berdasarkan IRT, estimasi skor sesungguhnya tergantung pada peluang menjawab benar pada kemampuan θ, karakteristik butir, dan model IRT yang digunakan.

Estimasi skor sesungguhnya, baik berdasarkan CTT maupun IRT, dapat dilakukan dengan berbagai model penskoran sehingga estimasi yang dihasilkan akan berbeda-beda. Oleh karena itu, kajian utama dari penelitian ini adalah menyelidiki implikasi penerapan model penskoran yang berbeda dalam mengestimasi skor sesungguhnya berdasarkan IRT. Model penskoran yang digunakan pada penelitian ini dibatasi hanya menggunakan model penskoran berdasarkan skor jumlah benar sesungguhnya (*number of right true score*), koreksi terhadap tebakan (*correction for guessing*), dan pembobotan optimal (*optimal weighting*).

Berdasarkan latar belakang masalah di atas, permasalahan yang akan dikaji dalam penelitian ini adalah: (1) karakteristik perangkat tes UN mata pelajaran matematika tahun pelajaran 2007/2008 tingkat SMP/MTs, (2) karakteristik distribusi skor sesungguhnya hasil estimasi dari beberapa model penskoran, (3) hubungan antara skor kemampuan (θ) dan skor tampak dengan skor sesungguhnya, dan (4) implikasi penerapan beberapa model penskoran terhadap perbedaan hasil estimasi skor sesungguhnya.

Kajian Pustaka

Pengukuran dilakukan untuk memperoleh data-data kuantitatif yang akan digunakan untuk menetapkan keputusan yang dikenakan kepada para siswa. Pengukuran didefinisikan sebagai suatu prosedur untuk memberikan nilai berupa angka yang disebut sebagai skor untuk spesifikasi atribut atau karakteristik seseorang di mana nilai itu menunjukkan derajat yang dimiliki oleh orang tersebut terhadap atribut yang sedang diukur. Jadi, pengukuran meliputi penetapan suatu nilai hasil suatu tes berdasarkan aturan tertentu. Berdasarkan karakteristiknya, pengukuran diklasifikasikan menjadi empat skala, yaitu skala nominal, ordinal, interval, dan rasio (Nitko & Brookhart, 2007: 7; Gronlund & Linn, 1990: 5)

Pengukuran terhadap suatu atribut, misalnya pengukuran terhadap kemampuan matematika, dapat dilakukan berdasarkan teori tes klasik dan teori respons butir. Teori respons butir memiliki sifat invariansi parameter, yaitu hasil pengukuran tidak tergantung pada kelompok siswa yang mengerjakan tes (*not group dependent*) dan hasil pengukuran tidak tergantung pada butir soal yang diujikan (*not item dependent*). Sementara teori tes klasik bersifat *group dependent* dan *item dependent* (Sotaridona, Pornel, & Vallejo, 2003: 84). Akibatnya, perbedaan estimasi kemampuan siswa terhadap karakteristik yang diukur tidak mencerminkan perbedaan kemampuan siswa (Hambleton, Swaminathan, & Rogers, 1991: 2).

Parameter butir dan parameter peserta yang bersifat invarians akan dicapai jika ada kecocokan data dengan model yang digunakan. Oleh karena itu, uji kecocokan model harus dilakukan karena jika model tidak cocok maka semua sifat invariansi IRT akan hilang sehingga keberfungsian IRT menjadi sia–sia (Welss, Hambleton, & Urip Purwono: 2008: 3). Taehoon Kang & Cohen (2007: 331) mengatakan bahwa pemilihan model IRT yang sesuai didasarkan pada kecocokan model dengan data dan derajat kompleksitas model. Model yang digunakan akan cocok jika data memenuhi asumsi-asumsi pada teori respons butir, yaitu unidimensi dan independen lokal (Hambleton & Swaminathan, 1985: 16; Hambleton, Swaminathan, & Rogers, 1991: 9).

Model yang digunakan untuk analisis tergantung pada tipe tes dan penskorannya. Akan tetapi, dalam praktik, pemilihan model yang cocok tergantung pada banyaknya data yang diperoleh untuk mengestimasi parameter–parameter dalam model yang digunakan. Secara umum, semakin banyak parameter yang diestimasi dalam suatu model, semakin banyak pula data (data respons siswa) yang dibutuhkan untuk mendapatkan hasil estimasi parameter yang baik (Stocking, 1999: 58-59). Ada tiga model logistik yang umum digunakan, yaitu model logistik 1 parameter, model logistik 2 parameter, dan model logistik 3 parameter.

Adapun model logistik 3 parameter ditunjukkan oleh persamaan (1) berikut:

$$P_i(\theta) = c_i + (1 - c_i)\,\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, 3, \ldots n \qquad (1)$$

dengan $P_i(\theta)$ adalah peluang menjawab benar butir *i* pada kemampuan θ, b_i adalah parameter tingkat kesulitan butir *i*, a_i disebut parameter pembeda butir *i*, c_i adalah parameter tebakan semu (*pseudo–guessing*), n adalah jumlah butir soal, $e = 2{,}718$, dan D = 1,7. Hambleton, Swaminathan, & Rogers (1991: 13) menyatakan bahwa nilai b_i yang baik bervariasi dari –2 sampai +2. Nilai parameter a_i terletak pada interval (0, 2) sedangkan nilai c_i dikatakan baik jika tidak lebih dari $1/k$, dengan k adalah banyaknya alternatif jawaban pada setiap butir soal (Hulin, Drasgow, & Parson, 1983: 36).

Skor sesungguhnya merupakan gabungan dari masing–masing peluang menjawab benar pada masing–masing butir soal yang diperoleh seorang siswa dengan kemampuan θ. Penggabungan peluang menjawab benar pada masing–masing butir soal pada kemampuan θ dapat dialakukan dengan berbagai rumus. Rumus-rumus ini kemudian disebut sebagai model penskoran (*formula scoring*).

Model penskoran yang pertama disebut model penskoran jumlah benar sesungguhnya (*number of right true score*). Estimasi skor sesungguhnya berdasarkan metode ini diperoleh dengan menjumlahkan peluang menjawab benar pada setiap butir soal dan menganggap setiap butir soal

memiliki bobot yang sama. Secara matematis dinyatakan sebagai berikut (Lord, 1980: 230; Baker, 2001: 66):

$$T_{NC} = \sum_{i=1}^{n} P_i(\theta) \quad (2)$$

dengan T_{NC} adalah skor jumlah benar sesungguhnya dari seorang peserta pada kemampuan θ, sedangkan n dan $P_i(\theta)$ seperti didefinisikan pada persamaan (1).

Model kedua adalah model penskoran koreksi terhadap tebakan (*correction for guessing*). Model penskoran ini mengoreksi jumlah skor benar sesungguhnya yang disebabkan oleh unsur tebakan dengan memberikan hukuman pengurangan skor pada soal yang dijawab salah karena jawaban yang salah dianggap sebagai jawaban hasil tebakan. Secara matematis model ini dinyatakan sebagai berikut:

$$T_{CG} = \sum_{i=1}^{n} P_i(\theta) - (\sum_{i=1}^{n} Q_i(\theta))/(k-1) \quad (3)$$

dengan T_{CG} adalah estimasi skor sesungguhnya berdasarkan koreksi terhadap tebakan, $Q_i(\theta) = 1 - P_i(\theta)$, dan k banyak pilihan jawaban.

Penerapan model penskoran koreksi terhadap tebakan mendapat kritik karena asumsi–asumsi yang digunakan sangat sulit terpenuhi dan secara umum gagal memenuhi maksud dari model tersebut, yaitu mengoreksi tebakan seperti dijelaskan oleh Thorndike (2005: 466-467). Model ini didasarkan pada asumsi bahwa semua tebakan bersifat buta. Asumsi ini menolak kemungkinan bahwa seorang siswa mungkin menebak berdasarkan pengetahuan parsial (*partial knowledge*) tentang butir soal yang sedang dihadapi. Berdasarkan kondisi ini seorang siswa dapat menghilangkan beberapa pilihan jawaban yang tidak mungkin atau tidak benar sehingga peluang menjawab benar karena menebak lebih besar dari $1/k$ (Allen & Yen, 1979: 167). Akibatnya, jika model penskoran koreksi terhadap tebakan diterapkan pada kondisi ini, maka skor terkoreksi yang diperoleh menjadi *underestimate*. Kritik lain adalah model ini tidak mengubah urutan peringkat siswa. Jadi, terdapat korelasi yang sempurna antara T_{NC} dan T_{CG}.

Model ketiga adalah model penskoran pembobotan optimum (*optimal weighting*). Pembobotan terhadap respons siswa dalam penelitian ini dilakukan secara implisit dengan besar bobot bervariasi berdasarkan model IRT yang digunakan (Rudner, 2000: 3). Masing–masing butir soal diberi bobot optimum yang berbeda sesuai dengan karakteristik butir soal dan berdasarkan model IRT yang digunakan. Secara umum skor komposit terbobot secara matematis dirumuskan sebagai berikut (Lord, 1980: 73):

$$T_{OW} = \sum_{i=1}^{n} w_i \, P_i(\theta) \qquad (4)$$

dengan w_i adalah nilai bobot butir i dan T_{OW} adalah estimasi skor sesungguhnya dengan pembobotan optimum (*optimal weighting*). Besarnya w_i ditentukan berdasarkan persamaan (5) sebagai berikut:

$$w_i(\theta) = \frac{Da_i}{1 + c_i \, e^{-D(\theta - b_i)}} \qquad (5)$$

Metode Penelitian

Pendekatan yang digunakan dalam penelitian ini adalah pendekatan deskriptif eksploratif. Data set dalam penelitian ini adalah respons siswa terhadap tes UN mata pelajaran matematika tingkat SMP/MTs tahun pelajaran 2007/2008 di Propinsi Nusa Tenggara Barat. Bentuk tes yang digunakan berupa tes pilihan ganda dengan empat pilihan jawaban dan terdiri dari 40 butir soal. Tes tersebut disusun dalam dua paket tanpa mengubah isi tes, tetapi hanya mengubah urutan nomor butir soal. Kedua paket soal tersebut adalah paket soal berkode P 61 dan P 28. Kedua paket soal ini diujikan pada kondisi yang sama sehingga faktor–faktor yang dapat mempengaruhi hasil tes di luar peserta tes relatif dikendalikan.

Data set yang dianalisis adalah data set respons siswa terhadap paket soal dengan kode P 28 sebanyak 1.624 orang. Data set diperoleh melalui teknik dokumentasi dengan cara mengutip respons siswa dari lembar jawab komputer. Adapun lembar jawab komputer tersebut didapatkan dari Dinas Pendidikan dan Olah Raga Propinsi Nusa Tenggara Barat.

Analisis data dimulai dengan mendeskripsikan kelayakan butir soal berdasarkan teori tes klasik dan program yang digunakan adalah program *ITEMAN* Versi 3.0. Sementara analisis berdasarkan teori respons butir menggunakan program *BILOG* Versi 3.07. Perbandingan skor sesungguhnya yang dihasilkan oleh masing–masing model penskoran dilakukan dengan memanfaatkan koefisien korelasi intraklas dan ANAVA pengukuran berulang (*repeated measure*) dengan bantuan SPSS Versi 15.

Hasil Penelitian dan Pembahasan

Hasil analisis karakteristik butir soal berdasarkan pendekatan teori tes klasik menunjukkan bahwa tingkat kesulitan butir soal mulai dari 0,259 sampai dengan 0,815 dengan rerata tingkat kesulitan butir soal sebesar 0,624. Menurut Allen & Yen (1979: 121), Gregory (2007: 153), dan Djemari Mardapi (2002: 116) tingkat kesulitan butir soal sebaiknya terletak pada interval 0,3 sampai 0,8 karena pada interval ini informasi tentang kemampuan siswa akan diperoleh secara maksimal. Berdasarkan kriteria tersebut, ada 2 butir soal (5%) dengan tingkat kesulitan soal rendah atau butir soal sulit dan 2 butir soal (5%) dengan tingkat kesulitan tinggi atau butir soal mudah.

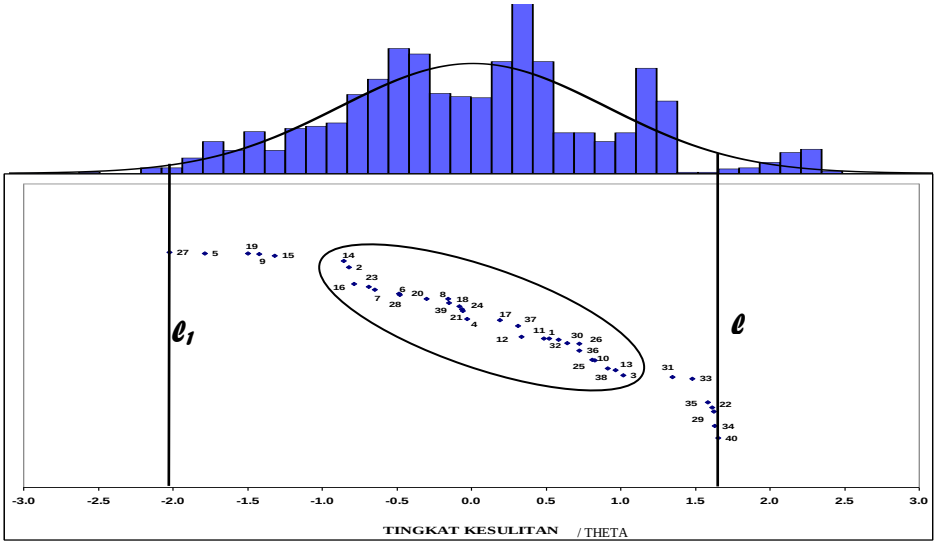
Jika penilaian yang digunakan adalah penilaian beracuan norma, maka berdasarkan rerata tingkat kesuliatan sebesar 0,624 dapat disimpulkan bahwa tes UN mata pelajaran matematika tahun pelajaran 2007/2008 sudah memenuhi syarat. Hal ini mengacu pada ketentuan yang diberikan oleh Gregory (2007: 154) bahwa rerata optimal untuk tes pilihan ganda dengan 4 pilihan jawaban sebesar 0,625. Perangkat tes dengan rerata sebesar 0,625 cocok untuk kelompok siswa yang berkemampuan sedang. Pada distribusi normal skor kelompok ini terletak pada $\pm 1\,SD$ (standar deviasi) atau terletak pada kelompok presentil ke 68%.

Berdasarkan daya pembeda butir, semua butir soal memiliki daya pembeda yang bernilai positif. Hal ini menunjukkan bahwa semua kunci berfungsi dengan baik. Setiap butir soal mampu membedakan siswa yang berkemampuan tinggi dengan siswa yang berkemampuan rendah. Siswa yang berkemampuan tinggi akan memilih kunci dan siswa yang

berkemampuan rendah memilih pengecoh. Tetapi, butir soal yang memiliki daya pembeda rendah masih dapat diterima jika penilaian yang digunakan adalah penilaian beracuan kriteria.

Analisis berdasarkan pendekatan IRT menunjukkan bahwa ada 6 butir soal (15%) yang kurang baik berdasarkan daya pembeda karena nilainya lebih dari 2,000. Berdasarkan tingkat kesulitan, secara umum butir–butir soal berada pada kriteria yang baik kecuali satu butir soal memiliki tingkat kesulitan terlalu rendah karena nilainya kurang dari –2,000. Sebanyak 25 butir soal (60%) yang kurang baik karena nilai *pseudo– guessing*–nya lebih dari 0,250 dan ada dua butir soal yang memiliki nilai *pseudo–guessing* yang sangat tinggi masing–masing sebesar 0,500 dan 0,480.

Gambar 1 menunjukkan pemetaan antara b dengan θ. Berdasarkan Gambar 1 diperoleh informasi bahwa semakin tinggi nilai b semakin sedikit prosentase banyaknya nilai θ yang lebih besar dari nilai b. Artinya, semakin sedikit jumlah siswa yang menjawab benar butir soal tersebut dengan peluang yang tinggi. Butir soal nomor 27 merupakan butir soal yang paling gampang $(b = -2.026)$ karena terdapat sekitar 99,32% siswa dengan nilai θ yang lebih besar dari nilai b sehingga memiliki peluang menjawab benar yang tinggi pada butir soal itu. Butir soal nomor 40 memiliki tingkat kesulitan paling tinggi $(b = 1,651)$ karena hanya 3,76% siswa yang memiliki peluang besar menjawab benar butir soal tersebut. Sementara butir–butir soal yang berada dalam daerah elips adalah butir–butir soal dengan nilai b mulai –0,854 sampai dengan 1,020. Butir–butir soal ini dapat dijawab dengan peluang benar tinggi oleh sekitar 16,01% sampai dengan 83,62% siswa.



Gambar 1. Pemetaan θ terhadap b

Tabel 1 merupakan statistik deskriptif skor yang diperoleh berdasarkan pendekatan CTT dan IRT yang meliputi estimasi skor sesungguhnya. Mencermati hasil pada Tabel 1 tampak bahwa rerata skor tampak (X) hampir sama dengan rerata skor sesungguhnya berdasarkan model penskoran jumlah benar sesungguhnya (T_{NC}) dan rerata kedua skor ini sedikit lebih tinggi dari rerata skor sesungguhnya berdasarkan model penskoran koreksi terhadap tebakan (T_{CG}) dan pembobotan optimum (T_{OW}). Jika dilihat berdasarkan nilai simpangan baku, penyebaran skor X, skor T_{NC}, dan skor T_{CG} relatif sama. Skor berdasarkan pembobotan optimum lebih menyebar dari reratanya.

Distribusi skor sesungguhnya berdasarkan ketiga model penskoran menunjukkan nilai *skewness* yang positif, artinya distribusi skor juling ke kanan yang menunjukkan bahwa sebagian besar siswa memperoleh skor yang rendah. Jika diperhatikan dengan seksama tampak bahwa skor T_{NC} dan T_{CG} menunjukkan nilai *skewness* dan *kurtosis* yang hampir sama. Hal ini

menunjukkan bahwa distribusi skor yang dihasilkan oleh kedua model tersebut tidak memiliki perbedaan yang signifikan. Bentuk distribusi skornya dapat dianggap berdistribusi normal karena nilai *skewness* dan *kurtosis* yang moderat, mendekati nol.

Tabel 1. Statistik Deskriptif Skor

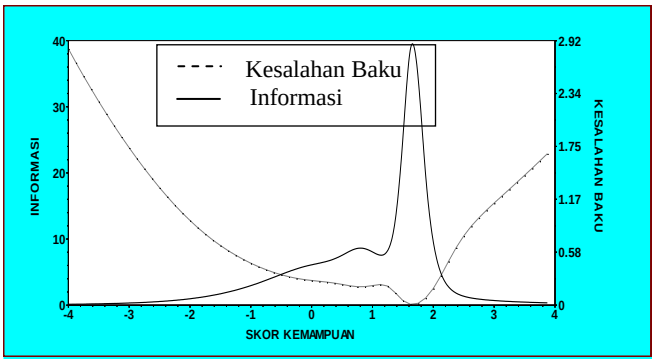
Statistik	Model Penskoran				
	X	θ	T_{NC}	T_{CG}	T_{OW}
Rerata	24,970	0,016	24,921	19,908	21,919
Simpangan Baku	5,624	0,922	4,739	6,321	14,634
Skewness	-0,382	0,122	0,370	0,375	1,973
Kurtosis	0,284	-0,070	-0,283	-0,281	5,132
Minimum	4	-2,694	15,170	7,670	2,550
Maksimum	40	2,465	37,070	36,100	78,440

Selanjutnya pada model T_{OW}, distribusi skor yang dihasilkan sangat menceng ke kanan karena nilai *skewness* yang sangat jauh dari nol dengan nilai kurtosis juga sangat tinggi. Hal ini menunjukkan bahwa berdasarkan model penskoran ini jumlah siswa yang mendapatkan skor rendah semakin banyak (skor di bawah rerata relatif homogen). Hal ini disebabkan oleh adanya pembobotan terhadap masing–masing butir soal yang nilainya sangat tergantung pada karakteristik butir soal dan nilai θ. Semakin rendah nilai θ, semakin rendah pula bobot butir soal sehingga skor sesungguhnya yang diperoleh rendah. Pada θ yang tinggi bobot skor optimal pada model 3P proporsional dengan daya pembeda butir, a_i.

Jika dicermati lebih jauh distribusi skor tampak dan distribusi skor sesungguhnya, khususnya berdasarkan model T_{NC} dan T_{CG}, memiliki nilai *skewness* dan *kurtosis* yang berbeda tanda. Distribusi skor tampak sedikit menceng ke kiri (nilai *skewness* negatif). Hal ini menunjukkan bahwa sebagian besar peserta mendapat skor yang tinggi atau di atas rerata. Sebaliknya, distribusi skor sesungguhnya juling ke kanan (nilai *skewness* positif) menunjukkan sebagian besar peserta mendapat skor yang rendah (di bawah rerata). Perbedaan ini terjadi karena pada penskoran dengan model skor tampak (teori tes klasik) pola respons siswa dan karakteristik

butir soal tidak diperhatikan sehingga memungkinkan beberapa orang siswa memiliki skor sama meskipun menjawab benar pada butir yang berbeda. Hal ini tidak berlaku pada estimasi skor sesungguhnya (teori respons butir) bahwa skor tampak yang sama tidak berarti memiliki estimasi kemampuan yang sama. Hal yang sama juga terjadi antara skor tampak dengan skor θ.

Gambar 2 memberikan informasi bahwa perangkat tes UN mata pelajaran matematika tahun pelajaran 2007/2008 hanya mampu mengukur kemampuan siswa pada kisaran interval [–0,450, 2,125]. Batas bawah dan batas atas interval tersebut merupakan skor kemampuan di mana grafik fungsi informasi dan grafik kesalahan baku pengukuran berpotongan. Pada interval tersebut kesalahan baku pengukuran yang terjadi kecil atau fungsi informasi yang diperoleh tinggi.



Gambar 2. Fungsi Informasi Tes dan Kesalahan Baku

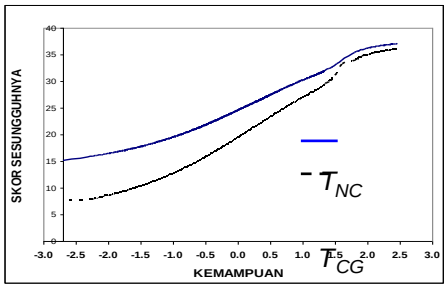
Berdasarkan hasil analisis fase III program BILOG Versi 3.07 dan Gambar 2 didapatkan bahwa informasi maksimum dicapai pada skor kemampuan 1,625 dengan nilai informasi 38,771 dan kesalahan baku pengukuran 0,129. Berdasarkan hasil ini dapat disimpulkan bahwa perangkat tes UN mata pelajaran matematika tahun pelajaran 2007/2008 tingkat SMP/MTs berdasarkan IRT cocok untuk kelompok siswa yang berkemampuan tinggi. Hal ini konsisten dengan hasil bahwa rerata tingkat kesulitan butir soal $(\bar{b} = 0{,}158)$ lebih tinggi dari rerata kemampuan

$(\bar{\theta} = 0,016)$ yang menunjukkan bahwa perangkat tes ini dalam kategori sulit.

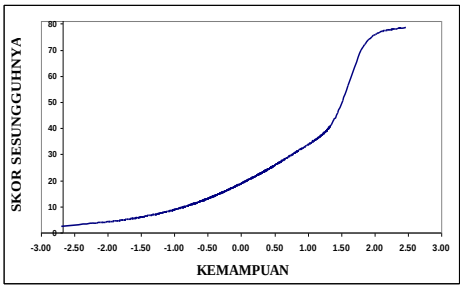
Informasi lain yang diperoleh dari Gambar 2 adalah kurva sangat curam pada θ di sekitar θ maksimum (1,625). Pada θ di sekitar θ maksimum diperoleh fungsi informasi yang sangat tinggi. Hal ini terjadi karena fungsi informasi berbanding lurus dengan kuadrat daya pembeda (a). Semakin tinggi nilai a semakin meningkatkan nilai fungsi informasi. Berdasarkan hasil analisis dengan BILOG Versi 3.07 diperoleh ada 5 butir soal dengan tingkat kesulitan hampir sama dengan θ maksimum $(b = 1,477 - 1,651)$ yang memiliki daya pembeda sangat tinggi $(a = 3,812 - 5,787)$.

Hubungan antara skor kemampuan dengan skor sesungguhnya dinyatakan oleh persamaan (2), (3), dan (4) yang lebih umum disebut sebagai kurva karakteristik tes (*test characteristic curve*). Kurva karakteristik tes merupakan hubungan fungsional antara skor sesungguhnya dengan skor kemampuan (Baker, 2001: 70). Gambar 3 dan 4 merupakan kurva karakteristik tes dari masing-masing skor sesungguhnya terhadap θ. Ketiga kurva tersebut menunjukkan bahwa ekor kurva paling bawah (asimptot) tidak sama dengan nol meskipun harga θ mendekati $-\infty$. Hal ini mencerminkan fakta bahwa siswa dengan kemampuan yang sangat rendah sekalipun bisa mendapatkan skor dengan hanya menebak sebagai akibat dari penerapan model 3P yang mengakomodir adanya faktor tebakan. Nilai statistik deskriptif masing–masing skor sesungguhnya dicantumkan pada Tabel 1.

Jika dicermati bentuk kurva karakteristik tes pada Gambar 3 hampir membentuk garis lurus sedangkan pada Gambar 4 bentuk kurva karakteristik tes adalah tidak linear pada bagian sebelah kanan. Secara umum, berdasarkan kurva karakteristik tes dan nilai korelasi masing-masing skor sesungguhnya terhadap θ (Tabel 2) menunjukkan bahwa semakin tinggi skor θ semakin tinggi pula skor sesungguhnya pada semua model penskoran.



Gambar 3: Kurva Karakteristik Tes Model T_{NC} dan T_{Cg}



Gambar 4 : Kurva Karakteristik Tes Model T_{OW}

Penerapan beberapa model penskoran untuk mengestimasi skor sesungguhnya menyebabkan hasil yang berbeda. Konsistensi atau kesesuain skor sesungguhnya antar masing-masing model penskoran ditunjukkan oleh besarnya nilai koefisien korelasi antar model penskoran seperti dirangkum pada Tabel 2. Koefisien korelasi dengan indeks P menyatakan koefisien korelasi *Product Moment Pearson* sedangkan koefisien korelasi dengan indeks A merupakan koefisien korelasi intraklas (*intraclass*) yang menyatakan kesesuaian skor.

Tabel 2. Koefisien Korelasi antar Model Penskoran

Model Penskoran	X	θ	T_{NC}	T_{CG}	T_{OW}
X	1,000$_P$	0,921$_P$	0,906$_P$ 0,892$_A$	0,905$_P$ 0,662$_A$	0,817$_P$ 0,527$_A$
θ		1,000$_P$	0,996$_P$	0,995$_P$	0,924$_P$
T_{NC}			1,000$_P$	1,000$_P$ 0,684$_A$	0,940$_P$ 0,531$_A$
T_{CG}				1,000$_P$	0,940$_P$ 0,674$_A$
T_{OW}					1,000$_P$

Mencermati hasil pada Tabel 2 tampak bahwa secara umum koefisien korelasi *Product Moment Pearson* antar model penskoran menunjukkan nilai yang sangat tinggi. Bahkan koefisien korelasi yang sempurna diperoleh

antara T_{NC} dan T_{CG} sehingga kedua model penskoran ini menghasilkan koefisien korelasi *Product Moment Pearson* yang relatif sama terhadap X, θ, dan T_{OW}. Hal ini menunjukkan bahwa penerapan model penskoran terhadap estimasi skor sesungguhnya tidak berimplikasi secara signifikan terhadap perubahan peringkat skor siswa.

Penerapan korelasi intraklas digunakan sebagai ukuran kesesuaian skor yang dihasilkan oleh masing-masing model penskoran (Prihoda, Pinckard, McMahan, at al, 2006: 379). Seperti dijelaskan sebelumnya bahwa koefisien korelasi *Product Moment Pearson* antara T_{NC} dan T_{CG} bernilai sempurna, namun tidak demikian dengan korelasi intraklas–nya. Hasil ini menunjukkan bahwa ketiga model penskoran menghasilkan skor yang memiliki urutan rangking yang sama, tetapi tidak menunjukkan kesesuaian skor.

Hasil pada Tabel 2 juga memberikan informasi bahwa koefisien korelasi intraklas antara X dengan T_{NC} paling tinggi dibandingkan dengan yang lain. Hal ini menunjukkan bahwa X paling sesuai dengan skor sesungguhnya yang dihasilkan dengan model T_{NC}. Tingkat kesesuaian skor antar model penskoran untuk estimasi skor sesungguhnya paling tinggi ditunjukkan oleh T_{NC} dan T_{CG} sedangkan tigkat kesesuaian skor antara T_{NC} dan T_{OW} lebih rendah dari tingkat kesesuain skor antara T_{CG} dan T_{OW}.

Berdasarkan hasil pada Tabel 3 diperoleh bahwa perbandingan antara skor tampak (X) dengan estimasi skor sesungguhnya menunjukkan perbedaan rerata yang tidak signifikan terjadi hanya antara X dengan T_{NC} karena memiliki *p-value* sebesar 1,000 yang lebih besar dari $\alpha = 0{,}05$. Hasil pada Tabel 1 menunjukkan bahwa kedua skor ini memiliki rerata yang hampir sama (24,970 dan 24,921). Berdasarkan hasil pada Tabel 2 diperoleh informasi bahwa koefisien korelasi intraklas paling tinggi terjadi antara X dengan T_{NC}. Hasil ini menjelaskan bahwa skor tampak (X) paling sesuai dengan skor sesungguhnya yang dihasilkan dengan model T_{NC} karena memiliki tingkat kesesuaian (*agreement*) skor paling tinggi.

Tabel 3. Uji Perbandingan Ganda

MODEL (I)	MODEL (J)	Rerata Perbedaan (I-J)	*Std. Error*	Sig.	Interval Konfidensi 95% untuk Perbedaan	
					Batas Atas	Batas bawah
X	T_{NC}	**0,053**	**0,060**	**1,000**	**-0,105**	**0,211**
	T_{CG}	5,065	0,067	0,000	4,889	5,241
	T_{OW}	3,055	0,262	0,000	2,363	3,746
T_{NC}	X	**-0,053**	**0,060**	**1,000**	**-0,211**	**0,105**
	T_{CG}	5,013	0,039	0,000	4,909	5,117
	T_{OW}	3,002	0,256	0,000	2,326	3,678
T_{CG}	X	-5,065	0,067	0,000	-5,241	-4,889
	T_{NC}	-5,013	0,039	0,000	-5,117	-4,909
	T_{OW}	-2,011	0,222	0,000	-2,598	-1,423
T_{OW}	X	-3,055	0,262	0,000	-3,746	-2,363
	T_{NC}	-3,002	0,256	0,000	-3,678	-2,326
	T_{CG}	2,011	0,222	0,000	1,423	2,598

Simpulan

Berdasarkan deskripsi hasil penelitian dan pembahasan yang telah dijelaskan dapat diambil kesimpulan sebagai berikut:

1. Berdasarkan pendekatan teori tes klasik, perangkat tes UN mata pelajaran matematika tahun pelajaran 2007/2008 tingkat SMP/MTs cocok diterapkan untuk penilaian yang beracuan norma dan cocok diberikan pada kelompok siswa yang memiliki kemampuan sedang atau pada kelompok siswa dengan kemampuan yang terletak pada presentil ke–68% pada distribusi normal. Sementara berdasarkan teori respons butir, perangkat tes tersebut hanya mampu mengukur kemampuan pada kisaran interval [–0,450, 2,125]. Rerata skor kemampuan diperoleh sebesar 0,016 lebih rendah dari rerata tingkat kesulitan sebesar 0,158. Hasil ini menunjukkan bahwa perangkat tes tersebut cocok untuk kelompok siswa yang berkemampuan tinggi. Hasil ini bertolak belakang dengan hasil berdasarkan teori tes klasik karena pada teori tes klasik tidak berlaku sifat invarians parameter sehingga karakteristik tes sangat

dipengaruhi oleh karakteristik siswa dan karakteristik siswa sangat ditentukan oleh karakateristik tes.

2. Penerapan model penskoran koreksi terhadap tebakan dan pembobotan optimum menyebabkan variansi skor semakin tinggi yang menunjukkan semakin heterogennya skor siswa yang akan berimplikasi pada jumlah siswa yang lolos dari batas *passing score* jika penilaian yang digunakan adalah penilaian beracuan kriteria.
3. Hubungan antara skor kemampuan (θ) dengan skor sesungguhnya dinyatakan melalui kurva karakteristik tes yang merupakan fungsi naik secara monoton. Skor kemampuan dan skor sesungguhnya berkorelasi positif dengan nilai yang sangat tinggi. Semakin tinggi skor kemampuan semakin tinggi pula skor sesungguhnya pada semua model penskoran. Artinya, skor sesungguhnya dapat menjadi pengganti skor kemampuan yang kurang familiar skala pengukurannya karena skor kemampuan dapat bernilai negatif sehingga menyulitkan dalam interpretasi.
4. Perbedaan rerata skor antara skor tampak dengan skor sesungguhnya berdasarkan model penskoran jumlah benar sesungguhnya tidak signifikan. Jika penilaian yang digunakan adalah penilaian yang beracuan norma, maka penerapan skor tampak dan skor sesungguhnya berdasarkan model penskoran jumlah benar sesungguhnya akan memberikan hasil yang relatif sama. Sebaliknya, jika penilaian yang digunakan adalah penilaian yang beracuan kriteria, skor sesungguhnya berdasarkan model penskoran jumlah benar sesungguhnya akan memberikan hasil yang lebih akurat. Secara umum penerapan skor tampak masih relevan jika penilaian yang digunakan adalah penilaian yang beracuan norma karena skor tampak dengan skor sesungguhnya dari ketiga model penskoran memiliki korelasi yang sangat tinggi.
5. Penerapan ketiga model penskoran terhadap estimasi skor sesungguhnya tidak berimplikasi pada perubahan peringkat skor sesungguhnya karena hasil estimasi ketiga model penskoran berkorelasi positif dan sangat tinggi satu sama lain. Namun, tingkat kesesuaian skor sesungguhnya dari masing–masing model penskoran relatif moderat (rendah). Implikasi dari hasil ini adalah jika penilaian yang digunakan merupakan penilaian yang beracuan kriteria, maka penerapan model

penskoran akan berimplikasi pada jumlah siswa lolos batas *passing score*. Model penskoran jumlah benar sesungguhnya akan memberikan jumlah siswa paling banyak yang melampaui batas *passing score*.

Saran

Isu tentang estimasi parameter butir dan kemampuan siswa berdasarkan IRT merupkan hal yang baru sehingga perlu diadakan seminar, *workshop*, dan diklat yang lebih banyak bagi para guru dan peneliti di bidang pengukuran. Pusat–pusat pengujian berbasis IRT perlu didirikan agar kegiatan pengukuran dan penilaian berbasis IRT dapat diterapkan pada ujian akhir semester, UN, UASBN, dan lain sebagainya.

Daftar Pustaka

Allen, M.J. & Yen, W.M. (1979). *Introduction to measurement theory*. Monterey: Wardsworth, Inc.

Baker, F.B. (2001). The basics of item response theory, 2nd Ed. *ERIC Clearinghouse on Assessment and Evaluation*. Diambil tanggl 17 Mei 2008 dari http://info.worldbank.org/etools/docs/library/117765/Item%20Response%20Theory%20-%20F%20Baker.pdf

Chopin, B. H. (1988). Correction for questing. (J. P. Keeves, ed.). *Educational Research, Methodology, and Measurement: An International Handbook* (pp. 384 – 386). Oxford: Pergamon Press.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Reinhart, and Winston, Inc.

Mardapi, D. (1999). *Estimasi kesalahan pengukuran dalam bidang pendidikan dan implikasinya pada ujian nasional*. Pidato pengukuhan guru besar, diucapkan di depan rapat senat terbuka UNY.

____. (2002). Bukti kesahihan dan keandalan alat ukur: Tanggapan atas artikel "Tes keterampilan olahraga judo bagi mahasiswa". *Jurnal Kependidikan, 1, 111 – 121*.

García–Pérez, M.A. & Frary, R.B. (1989) Psychometric properties of finite–state scores versus number–correct and formula scores: A simulation study. *Applied Psychological Measurement*, 13, 403–417. Diambil tanggal 25 Januari 2009 dari http://www.ucm.es/centros/cont/descargas/documento11676.pdf

Gregory, R.J. (2007). *Psychological testing. History, principles, and application, fifth edition.* New York: Pearson Education, Inc.

Gronlund, N.E. & Linn, R.L. (1990). *Measurement and evaluation in teaching*. New York: MacMillan Publishing Company.

Hambleton, R.K., Swamintahan, H., & Roger, H.J. (1991). *Fundamental of item response theory*. London: Sage Publication.

Hambleton, R.K., Swamintahan, H., & Roger, H.J. (1985). *Item response theory*. Boston: Kluwer Nijhoff Publishing.

Hulin, C.L., Dragsow, F., & Parson, C.K. (1983). *Item response theory application to psychological measurement*. Illionis: Dowjones-IRWIN.

Lord, F.M. (1980). *Application of item response theory to practice testing problem*. New Jersey: Lawrence Elbaum Associates.

Nitko, A., & Brookhart, S.M.(2007). *Educational assessment of students* (*5th Ed.*). New Jersey: PEARSON Merrill Prentice Hall.

Prihoda, T.J., Pinckard, R.N., McMahan, A., *et al.* (2006). Correction for guessing increases validity in multiple-choice examination in an oral and maxillofacial pathology course. *Journal of Dental Education*, 70, 378-386. Diambil tanggal 14 Juli 2008 dari http://www.jdentaled.org/cgi/reprint/70/4/378?ijkey=7ab362cad936711fc829e20a2ee6b7ceb3239b7e

Rudner, L.M. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice,* 20, 16-19

Sax, G. (1980). *Principles of educational and psychological measurement and evaluation*. Belmont: Wardsworth Publication Company, Inc.

Satoridona, L.S., van der Linden, W.J., & Meijer, R.R. (2006). Detecting answer copying using the Kappa statistic. *Applied Psycological Measurement*, 30, 412-431.

Simon, A.B., Budescu, D.V. & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, 21, 65-88. Diambil tanggal 14 Juli 2008 dari http://apm.sagepub.com/cgi/reprint/21/1/65

Stocking, M.L. (1999). Item response theory. Dalam G.N. Master & J.V. Keeves (Eds.), *Advences in Measurement in Educational Research and Assesment* (pp.43-54). Amsterdam: Pergamon.

Taehoon Kang & Cohen, A.S. (2007). IRT model selection methodes for dichotomous items. *Applied Psychological Measurement*, 31, 331-358. Diambil tanggal 25 Januari 2009 dari http://apm.sagepub.com/cgi/content/abstract/31/4/331

Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. Dalam D.Thissen & M. Orlando (Eds.), *Test Scoring* (73-140). London: Lawrence Erlbaum Associates.

Thorndike, R.M. (2005). *Measurement and evaluation in psychology and education, 7th edition*. New Jersey. Pearson Education, Inc.

Wary, J. (1995). *Critical value of questing on true-false and multiple choice tests*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Wells, C.S., Hambelton, R.K., & Urip Purwono. (2008). *Assessing the fit of IRT models to item response theory*, Makalah Disampaikan pada Pelatihan Psikometri, di Universitas Negeri Yogyakarta.