ANALISIS BUTIR DAN IDENTIFIKASI KETIDAKWAJARAN SKOR UJIAN AKHIR SEKOLAH UNTUK STANDARISASI PENILAIAN

Dadan Rosana dan Sukardiyono
Fakultas Matematika dan Ilmu Pengetahuan Alam UNY
email: danrosana@uny.ac.id

Abstrak

Penelitian ini bertujuan untuk mendapatkan hasil kajian empirik tentang kualitas butir soal pada UAS Mata Pelajaran Fisika di Kabupaten Lombok Timur tahun 2015 dengan menggunakan teori respon butir dan sekaligus untuk mengidentifikasi ketidakwajaran (*inappropriateness*) skor tes tersebut. Penelitian menggunakan metode deskriftif kuantitatif, ketidakwajaran skor dideteksi dengan korelasi person biserial (*person-fit statistic*) dan indeks kehati-hatian (*caution index*) dari Sato. Paket program yang digunakan untuk melakukan analisis butir adalah *Quest*, dengan elemen sentral *Rasch Model (RM)*. Subjek penelitian adalah 15 sekolah menengah atas sampel dari populasi 137 sekolah di Kabupaten Lombok Timur, yang diambil secara *purposive sampling*. Soal UAS digunakan bersama oleh 15 sekolah pada tahun ajaran 2014/2015. Hasil penelitian menunjukkan bahwa kualitas butir baik karena memiliki tingkat kecocokan dengan model Rasch dengan nilai *infit meansquare*-nya 0,77 -1,30. *Caution index* dari Sato terdeteksi 3,01 persen sampel mempunyai ketidakwajaran, dan 9,68 persen sampel yang diambil mempunyai tingkat ketidakwajaran yang tinggi.

Kata kunci: analisis butir, identifikasi ketidakwajaran, kualitas butir tes

THE ITEMS ANALYSIS AND THE IDENTIFICATION OF FINAL TEST SCORE INAPPROPIATENESS TO STANDARDIZE THE ASSESMENT

Abstract

This study was aimed at getting the result of an empirical study about the quality of items on Physics final exam held in Lombok Timur Regency 2015, using the theory of item response and identifying inappropriateness among those test scores. This study used descriptive quantitative method. The techniques used to identify the inappropriateness were the person-fit statistic and caution index as proposed by Sato. The program package used to hold the items analysis was Quest, by Rasch Model (RM) as the central element. The subjects were 15 high schools as the sample from the total of 137 high schools/MA in Lombok Timur Regency taken through purposive sampling. The items were used by the whole 15 sample schools for the final exam in the academic year of 2014/2015. The items analysis obtains good quality test items as it has a level of compatibility with Rasch models of its value infit mean square 0.77-1.30. The caution index proposed by Sato detected 3.01 % inappropriateness and the 9.68% high inappropriateness.

***Keywords**: items analysis, item test quality, the identification of inappropriateness*

PENDAHULUAN

Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2013 Tentang Kriteria Kelulusan Peserta Didik dari Satuan Pendidikan dan Penyelenggaraan Ujian Sekolah/Madrasah/ Pendidikan Kesetaraan dan Ujian Nasional, salah satu pasalnya mengungkapkan bahwa, nilai akhir (NA) untuk penentuan kelulusan diperoleh dari gabungan nilai

sekolah dari mata pelajaran yang diujikan secara nasional dan Nilai UN, yaitu dengan pembobotan 40% Nilai sekolah dari mata pelajaran yang diujikan secara nasional dan 60% dari Nilai UN.

Permasalahan yang kemudian muncul berkaitan dengan hal ini adalah belum adanya kesetaraan kualitas asesmen yang digunakan untuk penilaian di sekolah, padahal hasil ujian ini dijadikan bahan untuk penentuan kelulusan dan dijadikan data untuk pertimbangan penerimaan mahasiswa baru di Perguruan Tinggi. Hal ini tentu saja mengakibatkan tidak terpenuhinya rasa keadilan dari peserta didik dan calon mahasiswa baru, karena perbedaan kualitas tes yang diberikan mengakibatkan perbedaan nilai raport yang dijadikan dasar pengambilan keputusan kelulusan dari satuan pendidikan dan penerimaan mahasiswa baru di Perguruan Tinggi.

UAS dimaksudkan untuk mengukur kompetensi peserta didik yang ditetapkan dalam Standar Kompetensi Lulusan. Hasil yang diperoleh diharapkan benar-benar mampu menggambarkan kemampuan peserta didik. Seharusnya hasil yang diperoleh dapat membedakan peserta didik yang telah memenuhi dan yang tidak memenuhi standar yang ada pada standar kompetensi lulusan tersebut.

Namun ada kalanya skor peserta didik tidak sesuai dengan kemampuannya yang sebenarnya. Penyebabnya bisa diakibatkan oleh permasalahan yang muncul dari peserta didik dan bisa juga diakibatkan oleh kualitas butir tes yang diberikan pada mereka. Ada peserta didik yang keseharianya mempunyai prestasi belajar tergolong memadai ternyata memperoleh hasil pada UAS yang rendah dan akhirnya tidak lulus. Sebaliknya peserta didik yang keseharianya berprestasi rendah ternyata memperoleh hasil yang memuaskan pada ujian akhir sekolah.

Ketidakwajaran skor dapat terlihat dengan pola jawaban peserta ujian. Skor peserta ujian akan menjadi wajar apabila peserta tersebut dapat menjawab benar butir soal yang mudah dan menjawab salah butir soal yang sukar. Sebaliknya, skor peserta akan menjadi tidak wajar apabila peserta ujian menjawab salah soal yang mudah dan menjawab benar soal yang sulit (Naga, 2001: 43).

Kesalahan pengukuran yang diakibatkan peserta didik berikutnya terjadi ketika peserta didik mendapatkan skor lebih tinggi daripada kemampuan yang sebenarnya (*spuriously high*). Hal ini dapat terjadi ketika peserta didik mencontek atau memperoleh jawaban dari orang lain (Hulin, Drasgow, & Parsons, 1983: 17). Mereka akan memperoleh skor yang lebih tinggi dari kemampuan yang sebenarnya.

Hulin, Drasgow, & Parsons (1983: 11-112) mengemukakan bahwa pembahasan tentang ketidakwajaran pengukuran terbatas pada keanehan pola jawaban peserta tes (peserta didik) dalam tes. Bila pola jawaban yang dihasilkan peserta tes tidak normal. Misalnya, ada sejumlah jawaban benar terhadap butir-butir sulit pada seperdua tes yang pertama dan dan ada sejumlah jawaban salah terhadap butir-butir tes yang mudah pada seperdua tes berikutnya. Atau peserta tes yang kreatif mungkin memberikan penafsiran yang berbeda terhadap butir tes yang mudah. Akibatnya, respon butir seperti ini tidak cocok dengan teori respon butir yang mengasumsikan peluang jawaban benar sebagai fungsi dari kecerdasan (kemampuan) peserta tes.

Ketidakwajaran pengukuran dapat pula bersumber dari kondisi penilaian. Nitko (1996: 91-94) dan Wiersma & Jurs (1990: 340) menyatakan tekanan mental peserta tes, seperti cemas, khawatir, takut gagal, kekurangmampuan dalam menulis, dapat menyebabkan peserta tes tidak

berhasil menjawab secara benar butir-butir tes. Sebagai akibatnya peserta didik seperti ini akan memperoleh skor yang tidak tepat, yakni tidak sesuai dengan kemampuan mereka sebenarnya.

Penyebab lain dari ketidakwajaran pengukuran adalah kecurangan saat pelaksanaan tes, misalnya ada peserta didik yang menyontek. Pola jawaban dari peserta didik yang curang dalam tes mungkin akan tampak ganjil. Kelompok jawaban benar akan bercampur dengan kelompok jawaban yang hampir semuanya tidak benar. Jenis pola jawaban seperti ini berbeda pada setiap peserta ujian. Keadaan ini dapat dideteksi melalui analisis ketidakwajaran pengukuran.

Untuk mengatasi kesalahan hasil pengukuran yang diakibatkan peserta didik ini, terdapat beberapa indeks yang berdasarkan pola respon aktual (*actual observed response pattern*) seperti *caution index* yang dikembangkan oleh Sato, Indeks U yang dikembangkan oleh Van der Flier's, personal biserial yang dikembangkan oleh Donlon dan Fisher, dan *norm conormity index (NCI)* yang dikembangkan oleh Tatsuoka. Dalam penelitian ini digunakan teknik *personal biserial* dan *caution index*. Dalam pedoman baru untuk pengujian pendidikan yang adil disarankan untuk memeriksa validitas dari nilai tes individu melalui penggunaan *person-fit statistics* (Tendeiro & Meijer, 2014: 239).

Kesalahan pengukuran jenis kedua yang diakibatkan oleh kualitas butir tes, akan dianalisis menggunakan indeks yang didasarkan pada *item response theory*. Model pengukuran yang digunakan dalam penelitian ini adalah model model *likelihood* maksimum menggunakan model *logistic* satu parameter (L-1P) yang juga dikenal dengan sebutan *Rasch model*.

Rumus teoritis untuk perhitungan daya dan ukuran berdasarkan distribusi asimtotik estimator maksimum *likelihood* untuk model regresi logistik (Li, 2014: 441). Untuk itu digunakan program *Quest* terdiri dari suatu bahasa kontrol yang mudah digunakan dengan *output* yang informatif dan fleksibel.

Program *Quest* ini dapat digunakan untuk mengkonstruksi dan memvalidasi variabel data dikotomi (Raymond & Seik-Toon, 1996: 1) dan politomus beserta kombinasinya. Selain itu juga Program *Quest* ini dapat digunakan untuk melakukan analisis berdasarkan teori tes klasik (*Clasical Test Theory*).

Tujuan dilakukannya analisis butir soal adalah untuk meningkatkan kualitas soal, yaitu apakah suatu soal dapat diterima karena telah didukung data statistik yang memadai, diperbaiki karena terbukti terdapat beberapa kelemahan atau bahkan tidak digunakan sama sekali karena terbukti secara empiris tidak berfungsi sama sekali. Peningkatan kualitas tes dimaksudkan untuk menyesuaikan tingkat kesulitan butir tes dengan kemampuan peserta tes untuk tes dengan tujuan untuk mengetahui kemampuan siswa pada mata pelajaran tertentu (Mardapi, Haryanto, & Hadi, 2012: 131).

METODE

Populasi pada penelitian ini adalah seluruh soal dan lembar jawaban peserta pada UAS tingkat SMA/MA Mata Pelajaran Fisika seluruh Kabupaten Lombok Timur tahun pelajaran 2014/2015. Jumlah SMA di Lombok Timur yang terdaftar adalah 158 sekolah, terdiri dari SMA dan Madrasah Aliyah 137 sekolah dan SMK sebanyak 21 sekolah. Jumlah peserta didik di SMA/MA berjumlah 10142 orang dan SMK berjumlah 1269 orang.

Untuk penelitian ini, jumlah SMA yang dijadikan sampel adalah sebanyak 15 sekolah, dengan jumlah peserta didik

sebanyak 1363 orang. Sampel penelitian adalah soal dan lembar jawaban peserta didik yang diambil dari 15 sekolah yaitu SMA Negeri Labuhan Haji (N=113), SMA Negeri Selong (N=167), SMA Negeri 1 Suralaga (N=119), SMA 2 Selong (N=148), SMA 2 Sukamulia (N=48), SMA Negeri 1 Sakra (N=63), SMA Negeri 1 Jerowaru (N=26), SMA Negeri 1 Keruak (N=65), SMA Negeri 1 Terara (N=133), SMA Negeri 1 Pringsela (N=74), SMA Negeri 1 Montonggading (N=42), SMA Negeri 1 Sikur (N=110), SMA Negeri 1 Aikmel (N=103), SMA Negeri 2 Aikmel (N=71), dan SMA Negeri 1 Wanasaba (N=81).

Pemilihan sampel menggunakan teknik *purposive sampling*. Penentuan jumlah sampel dilakukan dengan menggunakan tabel penentuan jumlah sampel *Isaac* dan *Michael* yang termuat dalam Sugiyono (2012). Berdasarkan tabel, dengan tingkat kesalahan sebesar 5%, jumlah sampel yang digunakan pada penelitian ini adalah sebanyak 1363 lembar jawaban. Jumlah butir jangkar (*anchor*) yang digunakan dalam penelitian ini adalah 25%, yaitu 10 butir soal dari 40 butir soal UAS Fisika SMA di Lombok Timur tahun 2015. Hal ini mengacu pada pendapat Skaggs & Lissitz (Hayati & Mardapi, 2014: 31), yang menyatakan bahwa jumlah butir jangkar (*anchor*) yang digunakan minimal 20% dari jumlah butir soal.

Paket program yang digunakan untuk melakukan analisis butir dalam penelitian ini adalah Quest. Elemen sentral program *Quest* adalah *Rasch Model* (RM) satu parameter (1-PL). Program ini dapat menggunakan data respons yang diskor secara politomus. Program *Quest* dalam melakukan estimasi parameter, baik untuk item maupun untuk testi (*case*/*person*) menggunakan *unconditional* (UCON) atau *joint maximum likelihood* (Raymond & Siek-Toon, 1996: 89).

Model penskoran menggunakan model uji coba terpakai sehingga butir yang tidak *fit* dengan model dikeluarkan (tidak diperhitungkan) saat menentukan skor siswa. *Item Characteritic Curve* (ICC) akan mendatar (*flat*) bila besarnya INFIT MNSQ untuk butir atau *e* lebih besar dari satuan logit>1,30 akan berakibat membentuk *platokurtic curve*, jika satuan logit <0,77 akan terlalu runcing membentuk *leptokurtic curve* (Keeves & Alagumalai 1999: 36). Oleh karena itu, dalam program *Quest* ditetapkan bahwa suatu butir atau *person* dinyatakan *fit* dengan model dengan batas kisaran INFIT MNSQ dari 0,77 sampai 1,30 (Raymond & Siek-Toon, 1996: 30 & 90). Ada pula peneliti yang menggunakan batas yang lebih ketat, yakni dengan kisaran 0,83 sampai dengan 1,20 dan ada yang menggunakan pengujian berdasarkan besarnya nilai *INFIT t*, yakni menggunakan kisaran nilai *t* adalah ± 2 (pembulatan ± 1,96) jika taraf kesalahan/*alpha* sebesar 5% (Keeves & Alagumalai 1999: 34-36; Bond & Fox, 2007: 43).

Indeks ketidakwajaran skor berdasarkan teori respon butir yang digunakan berdasarkan hasil analisis program Quest di antaranya adalah ketidakwajaran melalui kebolehjadian dan indeks ketidakwajaran residu bakuan terkuadrat (Naga, 1992: 164-168).

Indeks kewajaran skor ditentukan melalui teori responsi butir. Karena estimasi parameter dilakukan melalui kebolehjadian maksimum, maka indeks kewajaran dihitung melalui kebolehjadian. Tingginya nilai kebolehjadian dijadikan indeks kewajaran; makin tinggi kebolehjadian makin wajar skor responden. Di dalam proses perhitungan digunakan logaritma, mencakup; indeks kewajaran l_0, indeks kewajaran l_g, dan indeks kewajaran l_z.

Indeks kewajaran kebolehjadian l_0 menggunakan logaritma dari kebolehjadian.

Kebolehjadian pada θ yang diestimasi melalui kebolehjadian maksimum.

$$L(X \mid \theta) = \prod_{i=1}^{N} P_i(\theta)^{X_i} Q_i(\theta)^{1-X_i} \quad (1)$$

dengan jawaban betul $X_i = 1$ dan jawaban salah $X_i = 0$.

Indeks kewajaran l_0

$$l_0 = \text{h } L(X \mid \theta)$$

$$= \sum^{N} \left[X_i \text{ h } P_i(\theta) + (1 - X_i) \text{ h } Q_i(\theta) \right] \quad (2)$$

dengan nilai $l_0 \leq 0$. Karena telah digunakan θ yang diperoleh melalui kebolehjadian maksimum, maka pada skor wajar seharusnya makin tinggi l_0 makin baik. Nilai l_0 yang rendah sekali menunjukkan ketidakwajaran skor. Jika butir mudah dijawab betul dan butir sukar dijawab salah, maka indeks kewajaran akan tinggi. Jika butir mudah dijawab salah dan butir sukar dijawab betul, maka indeks kewajaran akan rendah.

Indeks kewajaran kebolehjadian l_g mereratakan indeks kewajaran berdasarkan butir yang dijawab, sehingga menjadi:

$$l_g = e^{\frac{l_0}{N}} \quad (3)$$

dengan N = banyaknya butir yang dijawab. Karena perhitungan didasarkan pada indeks per butir yang dijawab, maka terdapat perlakuan sama di antara responden yang menjawab banyak butir dan yang sedikit butir. Makin tinggi nilai indeks kewajaran makin wajar skor responden.

Indeks kewajaran kebolenjadian nilai baku l_z. Apabila kemampuan responden θ berbeda, maka indeks kewajaran l_g menjadi kurang memadai. Untuk mengatasi hal ini, digunakan indeks kewajaran nilai baku.

$$l_z = \frac{l_0 - \mu_{l_0}}{\sigma_{l_0}} \quad (4)$$

Perhitungan indeks kewajaran memerlukan nilai rerata dan simpangan baku pada l_0.

$$\mu_{l_0} = \frac{\sum_{i=1}^{N} l_{0_i}}{N}$$

$$= \sum_{i=1}^{N} \left[P_i(\theta) \text{ h } P_i(\theta) + Q_i(\theta) \text{ h } Q_i(\theta) \right] \quad (5)$$

Melalui substitusi

$$m_i(\theta) = P_i(\theta) \text{ h } P_i(\theta) + Q_i(\theta) \text{ h } Q_i(\theta)$$

maka rerata menjadi

$$\mu_{l_0} = \sum_{i=1}^{N} m_i(\theta) \quad (6)$$

Simpangan bakunya adalah:

$$\sigma_{l_0} = \sqrt{\frac{N \sum_{i=1}^{N} l_{0_i}^2 - \left(\sum_{i=1}^{N} l_{0_i} \right)^2}{N^2}}$$

$$= \sqrt{\sum_{i=1}^{N} \left[P_i(\theta) Q_i(\theta) \left(\text{h } \frac{P_i(\theta)}{Q_i(\theta)} \right)^2 \right]} \quad (7)$$

Indeks Kewajaran Residu Bakuan Terkuadrat. Responden menghasilkan jawaban berupa jawaban betul $X_i=1$ dan jawaban salah $X_i=0$. Misalnya untuk model logistik menghasilkan probabilitas betul $P_i(\theta)$ dan probabilitas salah $Q_i(\theta)$. Selisih di antara mereka adalah residu R_i, yaitu:

$$R_i = X_i - P_i(\theta) \quad (8)$$

Residu menjadi dasar untuk menunjukkan kewajaran skor responden rerata dan simpangan baku sebagai berikut.

Rerata $\mu_{X_i} = P_i(\theta)$ (9)

Simpangan baku $\sigma_{X_i} = \sqrt{P_i(\theta) Q_i(\theta)}$ (10)

Nilai baku selisih atau residunya adalah:

$$S_{R_i} = \frac{X_i - \mu_{X_i}}{\sigma_{X_i}} = \frac{X_i - P_i(\theta)}{\sqrt{P_i(\theta)Q_i(\theta)}} \quad (11)$$

Pada saat $X_i=0$

$$S_{R_i}(X_i = 0) = -\sqrt{\frac{P_i(\theta)}{Q_i(\theta)}} \quad (12)$$

Pada saat $X_i=1$

$$S_{R_i}(X_i = 1) = \sqrt{\frac{Q_i(\theta)}{P_i(\theta)}} \quad (13)$$

Indeks kewajaran skor terkuadrat untuk *N* butir

$$W = \sum_{i=1}^{N} S_{R_i}^2 = \sum_{i=1}^{N} \left[X_i \frac{Q_i(\theta)}{P_i(\theta)} - (1 - X_i) \frac{P_i(\theta)}{Q_i(\theta)} \right] \quad (14)$$

Pada model logistik L-1P

$$\frac{Q_i(\theta)}{P_i(\theta)} = e^{-D(\theta - b_i)} \qquad \frac{P_i(\theta)}{Q_i(\theta)} = e^{D(\theta - b_i)} \quad (15)$$

Sehingga

$$W = \sum_{i=1}^{N} \left[X_i e^{-D(\theta - b_i)} - (1 - X_i) e^{D(\theta - b_i)} \right] \quad (16)$$

W diturunkan dari residu sehingga makin besar *W* makin besar residu dan makin tidak wajar skor responden.

Peluang peserta didik berhasil mengerjakan butir soal tergantung pada kemampuan (*ability*) dan tingkat kesulitan (*difficulty*) butir soal yang dikerjakannya (Isgiyanto, 2013:14). Untuk menggambarkan pola jawaban responden dihubungkan dengan kemampuan (*ability*) dan karakteristik butir soal seperti tingkat kesukaran digunakan *person-fit statistic*. *Person-fit statistic* dapat dijadikan sebagai indikasi kewajaran skor peserta ujian. Wright & Linacre (1992) mengungkapkan indeks ketidakwajaran dengan menggunakan *person-fit* dan menamakannya *weighted total fit mean square* dan *unweighted total fit mean square* (Rudner, 1983).

Person-fit statistic dapat diperoleh dengan melihat *outfit statistic* pada *output* Quest. *Outfit statistic person* menunjukkan bagaimana perilaku yang tidak diharapkan dari soal yang mempunyai tingkat kesukaran jauh dengan kemampuan peserta ujian tersebut (Setiadi, 1999). Dengan kata lain, peserta ujian dengan kemampuan tinggi tidak dapat menjawab soal dengan tingkat kesukaran rendah. Atau sebaliknya, peserta dengan kemampuan yang rendah tetapi menjawab benar butir soal dengan tingkat kesukaran tinggi. Hal ini sejalan dengan pengertian ketidakwajaran.

Untuk mempertajam analisis terkait indeks ketidakwajaran digunakan metode SHL diambil dari nama Sato, Harnisch dan Linn, yang selanjutnya disebut indeks kehati-hatian (*caution index*). Metode ini menggunakan banyaknya jawaban salah pada butir mudah dan banyaknya jawaban benar pada butir sulit.

Selanjutnya dicari indeks kehati-hatian dalam bentuk proporsi terhadap jawaban benar dari seluruh peserta. Dalam menentukan indeks tersebut, dibuat matriks jawaban peserta ujian. Karena fokus pada peserta ujian, maka pada bagian kolom pertama adalah peserta ujian dan pada bagian baris adalah butir soal. Selanjutnya, peserta diurutkan dari tinggi sampai rendah dimana peserta dengan tertinggi ditempatkan di atas. Butir soal juga diurutkan dari butir termudah sampai pada butir tertinggi.

Langkah selanjutnya adalah menentukan batas dari jawaban benar dan jawaban salah. Selanjutnya, yang perlu diperhatikan adalah pola jawaban peserta ujian. Berapa butir soal yang benar pada tempat yang seharusnya salah karena butir tersebut sukar. Juga berapa butir soal yang salah pada tempat yang seharusnya peserta tersebut benar karena butir tersebut mudah.

Beberapa notasi untuk rumus indeks kehati-hatian, adalah:

t = batas di antara jawaban salah dan jawaban betul jika responden berhati-hati

f_{gi} = skor butir pada indeks kehati-hatian untuk responden ke-g

f_t = banyaknya butir di bawah batas t

N = banyaknya butir

X_{gi} = skor butir oleh responden ke-g
= 1 untuk jawaban betul
= 0 untuk jawaban salah

Indeks kehati-hatian SHL untuk responden ke-g, dinyatakan dengan persamaan:

$$c_g = \frac{A_g - B_g}{C - D} \quad (17)$$

Dimana:

$$A_g = sekor\ \ jawaban\ \ salah$$

$$= \sum_{i=1}^{f_t} (1 - X_{\dot{g}})\, f_{\dot{g}} \quad (18)$$

$$B_g = sekor\ \ jawaban\ betul$$

$$= \sum_{i=f_t+1}^{N} X_{\dot{g}}\, f_{\dot{g}} \quad (19)$$

dan $C = \sum_{i=1}^{f_t} f_{gi}$ $\qquad D = \sum_{i=N-f_t+1}^{N} f_{gi}$ (20)

HASIL PENELITIAN DAN PEMBAHASAN

Data yang diperoleh dianalisis secara matematis dengan bantuan komputer menggunakan program *Quest*. Analisis program *Quest* yang dibahas dalam penelitian ini adalah bagian *output* menurut teori respon butir, yang terdiri dari; kecocokan dengan model, tingkat kesukaran butir tes, reliabilitas soal, dan estimasi kemampuan responden. Estimasi butir dan responden dilakukan dengan prosedur PROX (*normal approximation estimation*). Kecocokan antara kemampuan responden *(θ)* dan indeks kesukaran butir (b) akan menghasilkan akurasi dalam pengukuran. Akurasi maksimal terjadi saat *P (θ) = 0,5*. Estimasi parameter dilakukan dengan membuang responden yang benar dan salah semua. Estimasi parameter responden dan butir dilakukan serentak karena keduanya belum diketahui. Estimasi terus dilakukan sampai nilai parameter responden dan butir konstan.

Parameter pertama adalah kecocokan butir dengan model Rasch, yaitu dengan melihat nilai *infit meansquare*. Penetapan *fit* testi (*case/person*) secara keseluruhan dengan program *Quest* (Raymond & Siek-Toon, 1996) didasarkan pada besarnya nilai rata-rata INFIT *Mean of Square* (INFIT MNSQ) beserta simpangan bakunya. Penetapan *o* didasarkan pada besarnya nilai rata-rata INFIT *Mean of* INFIT *t*. Penetapan *fit* tiap testi (*case/person*) dengan model dalam program *Quest* didasarkan pada besarnya nilai INFIT MNSQ atau nilai INFIT *t* item yang bersangkutan. Langkah perhitungannya mengikuti langkah yang ditulis Wright & Masters (1982: 108-109).

Tahap kedua dalah melihat nilai *outfit t*. Tahap ketiga adalah menganalisis indeks kesukaran butir dengan melihat *thresholds*. Butir soal yang cocok dengan model Rasch memiliki *infit meansquare 0,77-1,30* nilai *outfit t* ≤2. Hampir semua soal dinyatakan cocok dengan model Rasch karena nilai *infit meansquare*-nya *0,77-1,30*. Soal nomor 14 dinyatakan gugur dengan nilai *outfit t*-nya *2,15*.

Parameter kedua yang dianalisis adalah tingkat kesukaran butir tes. Pada distribusi tingkat kesukaran soal dan kemampuan responden dapat dilihat pada file map. Bentangan tingkat kesukaran dan tingkat kemampuan siswa tersebut berada pada satu garis sehingga akan dapat diketahui posisi setiap subjek terhadap tingkat kesulitan butir yang dikerjakan. Tingkat kemampuan

tes maupun tingkat kesukaran butir dalam RM diekspresikan pada satu garis berupa absis pada grafik dengan satuan berupa logit (*logg-odd unit*) (Keeves & Alagumalai, 1999: 27).

Proporsi varians total dari estimasi skala untuk *person* sebesar *βn* yang berasosiasi dengan varians parameter, ditentukan oleh besarnya indeks separasi *person* sebesar *S*. Indeks separasi *person* itulah yang dianggap sama dengan koefisien reliabilitas tes (Wright & Masters, 1982: 106). Namun, tetap harus diperhatikan bahwa perhitungan besarnya *error* pengukuran pada indeks sparasi *person* berbeda dengan perhitungan *error* varians pada CTT (Keeves & Masters, 1999: 275-276).

File ini menyajikan pesebaran responden menurut tingkat kemampuannya dan persebaran butir menurut tingkat kesukarannya dalam logit *-2,0* sampai *+2,0*. Dari 40 butir soal yang dibuat, terdapat dua soal berkategori sangat mudah, delapan soal berkategori mudah, 25 soal berkategori sedang, empat soal berkategori sukar dan satu soal berkategori sangat sukar. Butir tes nomor sembilan dianggap soal yang paling mudah dan nomor 16 sebagai soal yang paling sukar.

Reliabilitas soal sebagai parameter yang ketiga mempunyai nilai 0,97 sehingga tes dianggap mempunyai reliabilitas sangat tinggi. Realibilitas ini terlihat dari *output* program *Quest* yang menyajikan hasil realiabilitas tes menurut CTT, yakni berupa indeks konsistensi internal, yang untuk penskor *an politomus* merupakan indeks alpha Cronbach dan untuk penskor *an dikotomus* merupakan indeks KR-20 (Raymond & Siek-Toon, 1996: 93). Ini berarti bahwa dengan program *Quest* dapat diperoleh item atau butir dan testi yang fit, disertai dengan reliabilitas instrumen tes tersebut.

Langkah selanjutnya adalah melakukan estimasi kemampuan responden. Kemampuan responden dapat dilihat dari banyaknya butir yang dapat dijawab dengan benar. Semakin banyak butir yang dapat dijawab dengan benar, maka kemampuan responden semakin tinggi. Estimasi kemampuan responden dapat dilihat pada file *teta* pada nilai *estimate*. Sampel pada penelitian ini sebanyak 1363 responden peserta UAS Mata Pelajaran Fisika SMA di Lombok Timur tahun 2015. Hasil analisis menunjukkan bahwa 254 responden (18,63%) memiliki kemampuan tinggi, 753 responden (55,25 %) memiliki kemampuan sedang, dan 356 responden (26,12%) memiliki kemampuan rendah.

Peta butir dan responden menggambarkan distribusi tingkat kesukaran butir dan kemampuan responden secara bersamaan dalam skala logit. Butir soal yang paling sukar yaitu butir item 23 yang berada di bagian teratas peta dengan tingkat kesukaran bernilai *2,27* sedangkan butir soal yang paling mudah adalah butir soal nomor 39 dengan tingkat kesukaran bernilai *-2,63*. Kemampuan resonden tertinggi terlihat pada tanda silang teratas yaitu bernilai *2,23* dan terendah *-2,18*. Dari analisis kualitas butir 40 soal pilihan ganda UAS Mata Pelajaran Fisika SMA di kabupaten Lombok Timur tahun pelajaran 2014/2015 secara kuantitatif baik menurut pendekatan teori respon butir.

Setelah dilakukan analisis butir berdasarkan data empirik hasil UAS Mata Pelajaran Fisika SMA di Kabupaten Lombok Timur tahun 2015, langkah selanjutnya adalah menganalisis ketidakwajaran (*inappropriateness*) skor tes tersebut. Ada dua teknik yang digunakan dalam menganalisis ketidakwajaran skor pada penelitian ini, yaitu teknik itu adalah *personal biserial* dan *caution index*.

Teknik *personal biserial* digunakan untuk melihat kesesuaian pola jawaban skor individu pada semua butir soal dan dibandingkan dengan pola jawaban skor kelompok. Pada teknik ini digunakan tingkat kesukaran yang dinyatakan dalam skala delta. Apabila pola jawaban individu sesuai dengan pola jawaban pada kelompok maka skor individu akan dinyatakan wajar. Sebaliknya, apabila pola skor yang diperoleh individu tidak sesuai dengan pola skor kelompok maka skor individu dinyatakan tidak wajar.

Nilai negatif pada indeks menunjukkan bahwa pada peserta ujian tersebut tidak terlihat adanya perbedaan kemampuan untuk menyelesaikan butir soal yang mudah dan sukar. Sebaliknya jika indeks *person biserial* semakin besar, maka pada peserta tersebut terlihat jelas perbedaan kemampuan untuk menyelesaikan butir soal yang mudah dan sukar. Apabila pada peserta ujian tidak terlihat jelas perbedaan kemampuannya antara butir mudah dan sukar, maka dapat dinyatakan bahwa terdapat ketidakwajaran skor pada peserta tersebut.

Hasil perhitungan, individu yang terdeteksi mempunyai skor tidak wajar pada SMA hanya 2,49% dari 1363 responden. Dari 15 sekolah yang menjadi sampel hanya empat sekolah yang terdeteksi mempunyai ketidakwajaran skor. Dari 34 responden yang terdeteksi mempunyai ketidakwajaran skor, 16 responden tersebut berasal dari satu sekolah sampel.

Berdasarkan perhitungan dengan teknik *caution index,* hanya 254 peserta (18,64%) yang dinyatakan mempunyai skor yang wajar. Untuk indeks rendah dan dapat ditolerir terdeteksi 662 peserta (48,57%) ujian. Selanjutnya 267 peserta (19,59%) terkategori mempunyai indeks ketidak wajaran sedang. Pada indeks kategori tinggi diperoleh angka 132 peserta (9,68%). Hanya 41 peserta (3,01%) yang dinyatakan mempunyai indeks yang sangat tidak wajar dan tujuh peserta (0,51%) mempunyai indeks negatif.

Analisis selanjutnya dengan menggunakan teknik *person-fit statistic* yang didasarkan pada *output Quest*. Analisis hasil UAS Mata Pelajaran Fisika SMA di Kabupaten Lombok Timur, diperoleh angka 2,78% responden (38 orang) mempunyai skor tidak wajar. Berdasarkan hasil tersebut, maka dapat dikatakan bahwa jumlah peserta yang mempunyai skor tidak wajar tergolong rendah.

Berdasarkan hasil analisis butir soal secara empiris menggunakan program *Quest* dari jawaban UAS Mata Pelajaran Fisika tahun 2015 di Lombok Timur, terlihat bahwa pada dasarnya telah diperoleh butir-butir tes yang cocok dengan model Rasch. Berdasarkan hal itu, akuntabilitas pengukuran telah dapat dipenuhi sebagaimana yang diharapkan. Hasil penelitian ini juga sangat memudahkan guru ketika ditugasi untuk membantu satuan pendidikan melakukan analisis butir soal untuk memenuhi bukti empiris di lapangan, sehingga dapat meningkatkan kualitas butir soal sampai menjadi soal yang terstandar.

Analisis butir soal dalam rangka mengembangkan soal UAS terstandar sangat dibutuhkan mengingat sudah banyak satuan pendidikan SMA yang membuat sendiri soal ujian sekolah. Teknik analisis inipun penting untuk digunakan dalam standarisasi soal tes yang dikembangkan oleh Dinas Pendidikan maupun MGMP sehingga dapat diketahui kualitas soal tersebut setelah dikerjakan oleh siswanya. Informasi dari kualitas butir soal tersebut menjadi masukkan dalam pengembangan perangkat butir soal UAS selanjutnya.

Berdasarkan temuan hasil analisis butir menggunakan *Quest*, maka di

lapangan yang berkait dengan penyusunan soal UAS yang terstandar, perlu direalisasikan upaya standarisasi tes berbasis sekolah atau MGMP. Satuan pendidikan harus mensikapi hal ini sebagai suatu bentuk otonomisasi dalam penilaian peserta didik sebagaimana dituntut di dalam Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2013 tentang Kriteria Kelulusan Peserta Didik dari Satuan Pendidikan dan Penyelenggaraan Ujian Sekolah/ Madrasah/Pendidikan Kesetaraan dan Ujian Nasional. Terlebih dengan adanya kebijakan penyelenggaraan penerimaan mahasiswa baru oleh perguruan tinggi negeri yang menggunakan Pangkalan data Sekolah yang berisi nilai perolehan siswa berdasarkan data hasil ujian oleh sekolah yang bersangkutan.

Dalam penggunaan program analisis butir seperti *Quest*, kesulitan memahami menu beserta pemanfaatannya ketika melakukan konversi merupakan permasalahan yang biasa terjadi di kalangan pendidik. Kondisi dapat diatasi dengan cara mengintensifkan pemanfaatan program analisis data hasil ujian untuk memperoleh bukti secara empiris, kalau perlu dicanangkan sebagai suatu tuntutan yang harus dipenuhi oleh setiap satuan pendidikan. Kebijakan yang mulai dilaksanakan pada tahun 2015, bahwa setiap satuan pendidikan wajib melaksanakan ujian sekolah untuk mata pelajaran yang diujikan secara nasional akan menjadi momen penting untuk dijadikan modal bagi satuan pendidikan untuk memberikan bukti empiris atas kualitas tes yang diujikan melalui ujian sekolah.

Diskusi selanjutnya adalah berkaitan dengan indeks ketidakwajaran soal, yang juga penting untuk bahan pembuatan kebijakan tentang pelaksanaan ujian sekolah. Hasil penelitian menunjukkan bahwa 3,01% sampel mempunyai ketidakwajaran, dan 9,68% sampel yang diambil mempunyai tingkat ketidakwajaran yang tinggi. Hal ini secara sederhana menunjukkan bahwa dengan teknik korelasi person biserial, skor peserta dinyatakan tidak wajar apabila memperoleh indeks negatif. Namun jika dibandingkan dengan korelasi *point biserial* atau korelasi *point biserial* untuk butir soal, maka selayaknya peserta yang mendapat indeks mendekati nol juga harus dinyatakan merupakan skor tidak wajar. Sebagai pembanding dapat kita lihat indeks daya beda butir soal yang dinyatakan oleh Ebel (dalam Crocker & Algina, 1986).

Untuk klasifikasi yang digunakan peneliti tentang *cuation index,* diperlukan analisis yang lebih mendalam dengan cara melakukan *replikasi*, sehingga dapat dilihat sejauh mana klasifikasi tersebut efektif. Naga (1992) menyatakan tentang pendapat para ahli yang menyatakan bahwa apabila indeks semakin besar maka skor tersebut semakin tidak wajar dan sebaliknya apabila semakin kecil maka skor tersebut semakin wajar.

Berdasarkan hasil analisis menggunakan *personal biserial* dan *caution index*, tentang ketidakwajaran skor terlihat bahwa antara grup satuan pendidikan tidak terjadi perbedaan yang signifikan. Namun pada ada beberapa sekolah, yang indeks ketidakwajaran sangat berbeda dengan sekolah yang lain. Ada sekolah yang respondennya mempunyai indeks ketidakwajaran yang cukup tinggi. Sebaliknya ada sekolah yang mempunyai peserta ujian relatif sedikit terindikasi ketidakwajaran skornya. Hal ini memperlihatkan bahwa kebijakan dan *fairness* setiap sekolah berbeda tentang pelaksanaan ujian sekolah.

SIMPULAN

Berdasarkan hasil penelitian dan pembahasan, dapat dikemukakan beberapa simpulan sebagai berikut ini. *Pertama*, telah didapatkan hasil kajian empirik tentang kualitas butir soal pada UAS Mata Pelajaran Fisika di Kabupaten Lombok Timur tahun 2015 dengan menggunakan teori respon butir, dan telah teridentifikasi ketidakwajaran (*inappropriateness*) skor tes tersebut. Hasil analisis program *Quest* yang telah diteliti adalah bagian output menurut teori respon butir, yang terdiri dari: kecocokan dengan model, tingkat kesukaran butir tes, reliabilitas soal, dan estimasi kemampuan responden. *Kedua*, sebagian besar soal dinyatakan cocok dengan model Rasch karena nilai *infit meansquare*-nya 0,77-1,30, kecuali soal nomor 14 dinyatakan gugur dengan nilai *outfit t*-nya 2,15. Dari 40 butir soal yang dibuat, terdapat 2 soal berkategori sangat mudah, 8 soal berkategori mudah, 25 soal berkategori sedang, 4 soal berkategori sukar dan 1 soal berkategori sangat sukar. Reliabilitas soal sebagai parameter yang ketiga mempunyai nilai 0,97 sehingga tes dianggap mempunyai reliabilitas sangat tinggi. *Ketiga*, hasil perhitungan dengan teknik caution index, 18,64 % mempunyai skor yang wajar, indeks rendah dan dapat ditolerir terdeteksi 48,57%, 19,59% terkategori mempunyai indeks ketidak wajaran sedang, indeks kategori tinggi 9,68%, dan 3,01% dinyatakan mempunyai indeks yang sangat tidak wajar. *Keempat*, hasil perhitungan dengan teknik *person-fit statistic* dari *output Quest*, terdapat 2,78% responden mempunyai skor tidak wajar. Hasil tersebut menunjukkan bahwa jumlah peserta yang mempunyai skor tidak wajar tergolong rendah.

DAFTAR PUSTAKA

Bond, T.G., & Fox, C.M. 2007. *Applying the Rasch Model: Fundamental Measurement in The Human Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Crocker, L., & Algina, J. 1986. *Introduction to Classical and Modern Test Theory*. New York: CBS College Publishing.

Hayati, N., & Mardapi, D. 2014. "Pengembangan Butir Soal Matematika SD di Kabupaten Lombok Timur sebagai Upaya dalam Pengadaan Bank Soal". *Jurnal Kependidikan*, 44(1), 26-38.

Hulin, C.L., Drasgow, F., & Parsons, C.K. 1983. *Item Response Theory: Application to Psychological Measurement*. Homewood, IL: Dow Jones-Irwin.

Isgiyanto, A. 2013. "Perbandingan Penyekoran Model Rasch dan Model Partial Credit pada Matematika". *Jurnal Kependidikan*, 43(1), 9-18.

Keeves, J., & Alagumalai, S. 1999. Advances in Measurement in Science Education. In Fraser, B. & Tobin, K. (Eds.), *International Handbook of Science Education*. Great Britain: Kluwer Academic Publishers.

Keeves, J.P., & Masters, G.N. 1999. Introduction. In Masters, G.N. & Keeves, J.P. (Eds), *Advences in Measurement in Educational Research and Assessment.* Amsterdam: Pergamon An Imprint of Elsevier Science.

Li, Z. 2014. "Power and Sample Size Calculations for Logistic Regression Tests for Differential Item Functioning". *Journal of Educational Measurement*. 51(4), 441-462.

Mardapi, D., Haryanto, & Hadi, S. 2012. "Pengujian Hasil Belajar dan Penilaian Pendidikan Berbantuan Komputer". *Jurnal Kependidikan*, 42(2), 130-143.

Naga, D.S. 1992. *Pengantar Teori Skor pada Pengukuran Pendidikan.* Jakarta: Penerbit Gunadarma.

Naga, D.S. 2001. "Indeks Kehati-hatian Skor Responden pada Model SHL: Suatu Bentuk Indeks Ketidakwajaran pada Skor Ujian". *Jurnal Ilmiah Psikologi "Arkhe"*, 6(1), April 2001.

Nitko, A.J. 1996. *Educational Assessment of Students*. (2nd eds). Columbus Ohio: Prentice Hall.

Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2013 tentang Kriteria Kelulusan Peserta Didik dari Satuan Pendidikan dan Penyelenggaraan Ujian Sekolah/Madrasah/Pendidikan Kesetaraan dan Ujian Nasional.

Raymond, A.J. & Siek-Toon, K. 1996. *Quest. The Interactive Test Analysus system*. The Australian Council for Educational Research.

Rudner, L.M. 1983. "Individual Assessment Accuracy". *Journal of Educational Measurement*, 20, 207-219.

Setiadi, H. 1999. "Penggunaan Program Bigstep untuk Pengembangan Bank Soal". *Makalah,* tidak diterbitkan.

Sugiyono. 2012. *Metode Penelitian Kombinsasi*. Bandung: Alfabeta.

Tendeiro, J.N., & Meijer, R.R. 2014. "Detection of Invalid Test Scores: The Usefulness of Simple Nonparametric Statistics". *Journal of Educational Measurement,* 51(3), 239-259.

Wiersma, W., & Jurs, S.G. 1990. *Educational Measurement and Testing.* (2nd ed.). Boston: Allyn and Bacon.

Wright, B., & Linacre, J. 1992. "Combining and Splitting Categories". *Rasch Measurement Transactions,* 6, 233-235.

Wright, B.D., & Masters, G.N. 1982. *Rating Scale Analysis*. Chicago: Mesa Press.