KARAKTERISTIK BUTIR SOAL UN BAHASA INGGRIS SMK DI KABUPATEN TABALONG KALIMANTAN SELATAN TAHUN AKADEMIK 2010/2011

Ahmad Hanafi [1)], Suhardi [2)]
SMK Negeri 1 Banua Lawas Tabalong Kalimantan Selatan [1)], Universitas Negeri Yogyakarta [2)]
hanafi08borneo@gmail.com [1)], suhardi@uny.ac.id [2)]

Abstrak

Hasil analisis kualitatif terhadap butir soal UN bahasa Inggris SMK TA 2010/2011 di Kabupaten Tabalong menunjukkan karakteristik bahwa dari 50 butir soal yang ada, setiap butir mengandung aspek materi, konstruksi, dan bahasa yang secara umum dapat dinyatakan sesuai. Kelemahan aspek materi, kurang sesuai standar kompetensi lulusan dan aspek konstruksi, pada penggunaan gambar, grafik, tabel, diagram atau sejenisnya yang belum berfungsi dengan baik. Hasil analisis kuantitatif menerapkan CTT mendapatkan temuan indeks kesukarannya baik, indeks daya pembeda, semua distraktornya berfungsi efektif, dan soal cukup reliabel. Analisis kuantitatif menerapkan IRT mendapatkan 27 butir cocok dengan model 3PL, meskipun sampel kurang memadai. Hasil pemetaan kualitas pembelajaran bahasa Inggris berbasis hasil UN bahasa Inggris SMK menunjukkan bahwa di Kabupaten Tabalong, SMK Tabalong paling unggul. Pada urutan berikutnya, SMKN 1 Tanjung, SMKN 1 Muara Uya, SMKN 1 Haruai, SMKN 1 Banua Lawas, dan SMK An Noor Paliat.

Kata Kunci: UN dan Karakteristik butir

THE CHARACTERISTICS OF THE ENGLISH TEST ITEMS IN NE FOR VHS'S IN TABALONG REGENCY SOUTH KALIMANTAN IN THE ACADEMIC YEAR OF 2010/2011

Abstract

The results of the qualitative analysis on the English test items in the NE for VHSs in Tabalong Regency in the academic year of 2010/2011 show that, of the 50 test items, each item contains the material, construction, and language aspects which are in general relevant. The weakness in the construction aspect is that pictures, graphs, tables, diagrams and the like do not function well. The results of the quantitative analysis using CTT reveal that the items are good based on the difficulty index, discrimination index, distractor effectiveness, and reliability. The quantitative analysis using IRT shows that 27 items fit IRT with the 3PL model, because the information is more complete and the standard error is smaller than the 1PL and 2 PL models although the sample is not adequate enough. The results of the mapping of the quality of English learning based on the results of the English test in the NE show that out of the six VHSs in Tabalong Regency, SMK Tabalong is at the top. The ranks below are occupied by SMKN 1 Tanjung, SMKN 1 Muara Uya, SMKN 1 Haruai, SMKN 1 Banua Lawas, and SMK An Noor Paliat.

Keywords: *NE, item characteristics*

PENDAHULUAN

Pembenahan mutu pendidikan di Indonesia secara nasional senantiasa dilakukan oleh pemerintah melalui kebijakan-kebijakan strategis Kemdikbud RI. Salah satu upaya yang dilakukan adalah menyelenggarakan evaluasi pendidikan secara nasional terhadap kompetensi peserta didik pada jenjang pendidikan dasar dan menengah. Salah satu bentuk evaluasi adalah menyelenggarakan Ujian nasional (UN) yang bersifat objektif, berkeadilan, dan akuntabel. Hasil UN sesuai ketetapan PP SNP pasal 68 ayat 1 akan digunakan sebagai salah satu pertimbangan pemetaan mutu program pendidikan secara nasional.

Sesuai kebijakan tersebut, di Kabupaten Tabalong Kalimantan Selatan telah dilakukan UN bahasa Inggris SMK TA 2010/2011. Tujuannya adalah sebagai salah satu syarat kelulusan dan memeta kualitas pembelajaran bahasa Inggris SMK secara nasional. Ujian Nasional pada dasarnya merupakan evaluasi pendidikan. Mardapi (2008, p.10) menyatakan bahwa ”hasil evaluasi yang bersifat nasional dapat dianalisis untuk memperoleh informasi yang akurat untuk perbaikan kualitas pendidikan”. Keberhasilan penyelenggaraan UN, salah satunya sangat bergantung pada karakteritik butir soal yang digunakan. Apabila karakteristik butir soal UN baik, akan mencerminkan bahwa kualitas tes yang digunakan juga baik dan hasilnya dapat dipertanggungjawabkan.

Karaktersitik butir soal bentuk tes pilihan ganda yang digunakan dalam UN dapat diketahui dengan melakukan analisis secara kualitatif pada aspek materi, konstruksi dan bahasa/budaya dan secara kuantitatif dengan teori tes klasik dan teori respons butir. Belum diketahuinya karakteristik butir soal UN bahasa Inggris SMK TA 2010/2011 di Kabupaten Tabalong berakibat pada hasil yang tidak dapat dipertanggungjawabkan. Apabila hasilnya tetap digunakan, maka penentuan kelulusan perlu ditinjau kembali dan peta kualitas pembelajaran bahasa Inggris SMK menjadi kurang akurat.

Ujian Nasional Bahasa Inggris SMK

Ujian Nasional (UN) merupakan kegiatan penilaian hasil belajar siswa yang telah meyelesaikan suatu jenjang pendidikan pada jalur sekolah/madrasah yang diselenggarakan secara nasional (Depdiknas, 2009, p.29). UN berdasarkan ketetapan pasal 66 ayat 1 Peraturan Pemerintah RI Nomor 19 tahun 2005 tentang SNP, didifinisikan sebagai penilaian hasil belajar yang dilakukan oleh pemerintah dengan tujuan untuk menilai pencapaian kompetensi lulusan secara nasional pada mata pelajaran tertentu dalam kelompok mata pelajaran ilmu pengetahuan dan teknologi.

Sekolah kejuruan memiliki tujuan khusus. Premono (2010, p.52) menegaskan bahwa SMK sebagai salah satu institusi yang menyiapkan tenaga kerja dituntut mampu menghasilkan lulusan sebagaimana yang diharapkan dunia kerja. Permendiknas RI Nomor 23 tahun 2006 tentang Standar Kompetensi Lulusan, menetapkan tujuan SMK adalah meningkatkan kecerdasan, pengetahuan, kepribadian, akhlak mulia, serta keterampilan untuk hidup mandiri dan mengikuti pendidikan lebih lanjut sesuai dengan kejuruannya.

Bahasa Inggris sebagai salah satu mata pelajaran wajib bagi siswa SMK juga didesain khusus sesuai kebutuhan yang dikenal dengan *English for Specific Purpose* (ESP). *English for Specific Purpose* berfungsi membantu siswa mengatasi penguasaan bahasa untuk kebutuhan pengembangan kompetensi sesuai disiplin dan profesi kerja (Basturkman, 2006, p.6). Hutchinson & Water (1991, p.21), berpendapat bahwa ESP merupakan suatu pendekatan untuk pembelajaran bahasa yang mana segala keputusannya seperti konten dan metodenya didasari atas tujuan pelajar untuk belajar. Sieroeka (2013, p.33), bahwa ESP dikendalikan berdasarkan kebutuhan untuk menggunakan bahasa sebagai alat dalam memfasilitasi kesuksesan dalam kehidupan profesional. Siswa diharapkan mampu mempresentasikan bidang ilmu yang menjadi konsentrasi jurusan sekolahnya, sehingga metode pembelajaran bahasa yang dipergunakan, sesuai standar yang ditetapkan pihak yang berkepentingan.

Kelompok mata pelajaran yang berisi deskripsi kelompok mata pelajaran spesifik SMK, merujuk pada Permendiknas RI Nomor 22 Tahun 2006, meliputi tiga kelompok mata pelajaran, yaitu kelompok normatif, kelompok adaftif, dan kelompok produktif. Mata pelajaran bahasa Inggris itu sendiri di SMK termasuk dalam kelompok adaptif. Berdasarkan lampiran salinan Permendiknas Nomor 46 tahun 2010 tentang kisi-kisi UN mata pelajaran bahasa Inggris SMK yang terdiri atas dua SKL, seperti tampak pada Tabel 1 berikut ini.

Tabel 1. Kisi-kisi UN Bahasa Inggris SMK TA 2010/2011

SKL	Indikator	Nomor Butir
Mendengarkan (*listening*)	*a. Picture*	1 – 4
	b. Questions response	5 – 8
	c. Short conversation	9 – 11
	d. Short talk (monolog)	12 – 15
Membaca (*reading)*	*a. Incomplete dialog*	16 – 32
	b. Error recognition	31 – 36
	c. Reading comprehension	37 – 50

Sesuai dengan tabel kisi-kisi UN bahasa Inggris bagi Siswa SMK TA 2010/2011 hanya memuat dua standar kompetensi lulusan (SKL), yaitu mendengarkan dan membaca. Apabila ditinjau lebih jauh, kompetensi bahasa Inggris selain mendengarkan dan membaca juga ada kompetensi atau keterampilan berbicara dan menulis (Tarigan, 2008, p.2). Hal ini berarti bahwa UN bahasa Inggris bagi Siswa SMK TA 2010/2011 belum menguji seluruh komponen kemampuan berbahasa Inggris.

Telaah Butir Soal UN

Analisis tes pendekatan kualitatif dilakukan dengan penelaahan tes secara teoritis sebelum tes diujicobakan, sehingga banyak yang menyebut teknik ini sebagai analisis tes secara teoritis. Telaah tes secara teoritis ini didasarkan pada kaidah penulisan soal yang menurut Mardapi (2008, p.137), meliputi aspek materi, konstruksi, dan bahasa. Zulaiha (2008, p.1), menjelaskan analisis tes secara teoritis ini dilakukan berdasarkan pertimbangan ahli atau profesional judgement. Hal ini dilakukan untuk meyakinkan bahwa tes yang dikembangkan berkualitas baik. Berikut ini pedoman telaah tes bentuk pilihan ganda berdasarkan aspek materi, konstruksi, dan bahasa.

Aspek materi meliputi beberapa komponen yang harus ditelaah, di antaranya soal harus sesuai dengan indikator, pilihan jawaban harus homogen dan logis ditinjau dari segi materi, dan setiap soal harus mempunyai satu jawaban yang benar atau paling benar Zulaiha (2008, p.2). Depdiknas (2010, p.125) menambahkan bahwa materi harus memiliki relevansi dengan kompetensi.

Aspek konstruksi meliputi beberapa komponen yang harus dicermati atau ditelaah, diantaranya pokok soal dirumuskan secara jelas dan tegas, memuat pernyataan yang diperlukan saja, tidak memberikan petunjuk ke arah jawaban, tidak mengandung pernyataan negatif ganda, panjang pilihan jawaban relatif sama, tidak mengandung pernyataan semua pilihan jawaban benar atau semua salah, dan pilihan jawaban disusun menurut besar kecil atau kronologi. Selanjutnya, gambar (termasuk grafik, tabel, diagram, dan sejenisnya) harus jelas dan berfungsi, dan butir jangan bergantung pada jawaban sebelumnya (Zulaiha, 2008, p.2; Depdiknas, 2010, p.125).

Aspek bahasa meliputi beberapa komponen yang harus ditelaah, diantaranya menggunakan kaidah bahasa yang benar dan tepat, jangan menggunakan bahasa tabu (berlaku di daerah setempat), menggunakan bahasa yang komunikatif, dan tidak ada pengulangan kata pada pilihan jawaban (Zulaiha, 2008, pp.2-3; Mardapi, 2008, p.157).

Teori Tes Klasik

Skor dari hasil suatu pengukuran yang belum dioleh (belum diderivasikan) disebut sebagai skor perolehan (*obtained–scores* atau *observed–scores*) dan dapat pula dinamakan skor tampak (simbol X_{ij}). Adanya skor tampak, berarti terdapat skor tidak tampak yang merupakan angka sebagai harga yang merepresentasikan secara murni performasi individu dengan benar. Oleh karena tidak dapat diungkap secara langsung oleh alat ukur (tes), sehingga tidak pernah dapat diketahui secara pasti besaranya. Skor yang merepresentasikan secara murni performasi individu dengan benar ini disebut skor sesungguhnya (*true scores*) atau skor murni (simbol T_{ij}). Selain adanya skor tampak dan skor murni yang diungkap, setiap hasil pengukuran selalu disertai adanya komponen kesalahan atau *error* (simbol E_{ij}). Besarnya komponen kesalahan tersebut bagi setiap individu dalam setiap tes juga tidak dapat diketahui secara pasti.

Berbagai Asumsi CTT

Adanya skor tampak X_{ij}, skor murni T_{ij}, dan komponen *error* E_{ij}, maka keterkaitan di antara ketiganya dapat diuraikan dalam beberapa asumsi yang merupakan asumsi teori skor–murni klasik (*classical true–scores theory*) berikut ini Allen & Yen, 1979, pp.57-59), (a) asumsi 1: $X = T + E$. Besarnya skor tampak (X) ditentukan bersama oleh besarnya skor murni (T) dan besarnya eror pengukuran (E), (b) asumsi 2: $e(X) = T$. Skor murni (T) merupakan nilai harapan X $(e(X))$. Skor murni merupakan harga rata-rata dari distribusi teoritik skor tampak apabila orang yang sama dikenai tes yang sama berulangkali dengan asumsi

pengulangan tes itu tidak terbatas banyaknya sedangkan setiap pengulangan tes adalah independen satu sama lain, (c) asumsi 3: $\rho_{et} = 0$. Bagi suatu kelompok populasi subjek yang dikenai tes, distribusi eror pengukuran tidak berhubungan satu sama lain. Ragam eror tidak berkaitan dengan ragam skor murni, (d) asumsi 4: $\rho_{e1e2} = 0$. Besarnya eror pada suatu tes tidak bergantung pada eror tes lain. Seorang subjek yang skornya pada tes pertama mengandung eror besar, tidak berarti akan mempunyai eror yang besar pula pada tes kedua, (e) asumsi 5: $\rho_{e1t2} = 0$. Eror pada suatu tes tidak bergantung kepada skor murni tes lainya. Meski demikian, asumsi ini tidak akan bertahan apabila salah satu tes yang bersangkutan ternyata mengukur aspek yang berpengaruh terhadap terjadinya eror pada pengukuran lainnya.

Indeks Kesukaran

Indeks kesukaran butir sebagaimana dinyatakan oleh Allen & Yen (1979, p.120) adalah *proportion of examinees who get that item correct.* indeks kesukaran adalah proporsi peserta ujian yang menjawab benar. Hampir serupa, Zaman et al (2010, p.62) menyatakan bahwa indeks kesukaran butir adalah persentase atau rasio siswa yang menjawab butir soal dengan benar pada suatu butir.

Proporsi menjawab benar *p* (*proportion correct*) adalah indeks kesukaran soal yang paling sederhana dan sering digunakan dalam menentukan besaran indeks. Rumus untuk menentukan besarnya indeks kesukaran secara matematis dirumuskan oleh Mardapi (2008, p.134) yang kemudian diadaptasi menjadi sebagai berikut.

$$p_i = \frac{n_i}{N}$$

Indeks kesukaran butir diberi simbol p_i, simbol n_i adalah jumlah peserta tes yang menjawab benar, sedangkan simbol N adalah banyaknya siswa yang menjawab butir soal tersebut. Besarnya indeks korelasi berkisar antara 0 sampai 1. Semakin tinggi besaran indeks kesukaran, maka butir soal dapat dinyatakan semakin mudah. Allen & Yen (1979, p.121) berpendapat bahwa tingkat kesukaran yang baik adalah 0,3 sampai 0,7. Butir dengan tingkat kesulitan di bawah 0,3 dianggap butir soal yang terlalu sukar, sedangkan jika indeksnya di atas 0,7 butir soal tersebut dianggap terlalu mudah.

Indeks Daya Pembeda

Zaman et al (2010, p.62) menjelaskan bahwa indeks diskriminan atau daya pembeda merupakan kemampuan siswa peserta tes dalam menjawab suatu tes hubungannya dengan total tes. Daya beda (diskriminan) suatu butir tes adalah kemampuan suatu butir untuk membedakan antara peserta tes yang berkemampuan tinggi dan berkemampuan rendah. Penentuan daya beda butir biasanya dilakukan dengan menggunakan indeks korelasi, diskriminasi, dan indeks keselarasan item. Dari ketiga cara tersebut yang paling sering digunakan adalah indeks korelasi. Ada empat macam teknik korelasi yang biasa digunakan untuk menghitung daya beda, menurut Naga (1992, pp.67-107) yaitu: (1) teknik *point biserial*, (2) teknik *biserial*, (3) teknik phi, dan (4) teknik tetrachorik.

Daya beda juga dapat dijelaskan sebagai derajad hubungan antara skor butir dengan skor total dengan menggunakan teknik korelasi *product moment* dari Pearson. Rumus korelasi *product moment* cocok untuk data politomi, sedangkan untuk data dikotomi yang cocok adalah korelasi *point biserial* seperti rumus berikut ini (Chada, 2009: 110).

$$r_{pbis} = \frac{Mp_i - Mq_i}{s_X}\sqrt{p_i q_i}$$

Di mana Mp_i, mean total skor peserta yang memiliki jawaban benar. Mq_i adalah mean skor total s_X, adalah standar deviasi skor total, p_i adalah proporsi peserta ujian yang menjawab benar pada butir tes sedangkan q_i adalah $1- p_i$. Rumus korelasi *point biserial* juga dapat diturunkan langsung dari rumus korelasi produk momen tanpa membuat pembatasan asumsi.

Efektivitas Distraktor

Setiap tes pilihan ganda memiliki satu pertanyaan serta beberapa pilihan jawaban. Diantara pilihan jawaban yang ada, hanya satu yang benar. Selain jawaban yang benar tersebut, adalah jawaban yang salah. Jawaban yang salah itulah yang dikenal dengan *distractor* (pengecoh). Dengan demikian, efektifitas distraktor adalah seberapa baik pilihan yang salah tersebut dapat mengecoh peserta tes yang memang tidak mengetahui kunci jawaban yang tersedia. Semakin banyak peserta tes yang memilih distraktor tersebut, maka distaktor itu dapat menjalankan fungsinya dengan baik.

Cara menganalisis fungsi distraktor dapat dilakukan dengan menganalisis pola penyebaran jawaban butir. Pola penyebaran jawaban sebagaimana dikatakan Zulaiha (2008, p.17) adalah diperolehnya penyebaran jawaban akan diketahui berfungsi tidaknya pengecoh. Dengan kata lain penyebaran jawaban merupakan suatu pola yang dapat menggambarkan bagaimana peserta tes dapat menentukan pilihan jawabannya terhadap kemungkinan jawaban yang telah dipasangkan pada setiap butir.

Menurut Mardapi (2008, p.159) sebuah distraktor dapat dikatakan berfungsi dengan baik jika dipilih oleh paling sedikit 5%. Sedangkan menurut Zulaiha (2008, p.17) distraktor dikatakan baik jika dipilih oleh minimal 2,5% dari seluruh peserta. Distraktor yang tidak memenuhi kriteria tersebut sebaiknya diganti dengan distraktor lain yang mungkin lebih menarik minat peserta tes untuk memilihnya.

Validitas dan Reliabilitas

Validitas seringkali dikonsepkan sebagai sejauhmana tes mampu mengukur atribut yang seharusnya diukur. Sesuai CTT, pengertian validitas dinyatakan dengan sejauhmana skor tampak dapat mendekati besarnya skor murni. Suatu alat ukur yang tinggi validitasnya akan menghasilkan error pengukuran yang kecil, artinya skor setiap subjek yang diperoleh oleh alat ukur tersebut tidak jauh berbeda dari skor yang sesungguhnya. Di dalam CTT, konsep validitas dinyatakan sebagai validitas instrinsik, yang dirumuskan sebagai akar kuadrat dari rasio antara varian skor murni dan varian skor tampak.

Sebuah tes dikatakan valid jika itu memang mengukur apa yang seharusnya diukur (Allen & Yen, 1979, p.95). Dalam bahasa yang hampir sama Mardapi (2008, p.25) menyatakan bahwa validitas adalah ukuran seberapa cermat suatu tes melakukan fungsi ukurnya. Menurut Kaplan & Saccuzzo (2005, p.38) kevalidan sebuah alat ukur tergantung pada konsep kesatuan yang mewakili semua bukti yang mendukung penafsiran yang dimaksudkan dari pengukuran.

Dari cara mengestimasinya yang disesuaikan dengan sifat dan fungsi setiap tes, tipe validitas umumnya digolongkan ke dalam tiga kategori, yaitu validitas isi, validitas konstruk, dan validitas berdasarkan kriteria. Validitas isi dibagi menjadi dua tipe, yaitu validitas muka dan validitas logik. Validitas konstruk adalah tipe validitas yang menunjukkan sejauh mana tes mengungkap suatu *trait* atau konstruk teoritik yang hendak diukurnya. Prosedur validasi berdasarkan kriteria menghasilkan dua macam validitas, yaitu: prediktif dan konkuren.

Di dalam suatu alat ukur atau tes, semakin besar porsi varian *error*, maka akan semkian kurang handal, dan tes akan semakin handal apabila memiliki semakin kecil porsi varian *error*. Kehandalan dalam hal ini juga dapat disebut reliabilitas alat ukur atau tes. Berdasarkan asumsi CTT, dapat diuraikan enam cara mengintepretasikan koefisien reliabilitas tes (Azwar (2010, pp.33-37), (a) intepretasi 1: $\rho_{XX'}$ = koefisien reliabilitas merupakan korelasi skor–tampak antara dua tes yang paralel, (b) interpretasi 2: $\rho_{XX'}^2$= koefisien reliabilitas merupakan besarnya proporsi varian *X* yang dijelaskan oleh korelasi linear *X*, (c) *i*ntepretasi 3: $\rho_{XX'} = \sigma_T^2 / \sigma_X^2$, koefisien reliabilitas merupakan perbandingan antara varian skor murni dan varian skor tampak pada suatu tes, (d) intepretasi 4: $\rho_{XX'} = \rho_{XX'}^2$, koefisien reliabilitas merupakan kuadrat kuefisien korelasi antara skor tampak dengan skor murni, (e) intepretasi 5: $\rho_{XX'} = 1 - \rho_{XE}^2$, koefisien reliabilitas merupakan Satu dikurangi kuadrat koefisien korelasi antara skor tampak dengan eror pengukuran, (f) intepretasi 6: $\rho_{XX'} = 1 - \sigma_E^2 / \sigma_T^2$, koefisien reliabilitas merupakan proporsi varian eror yang dicerminkan oleh varian skor tampak.

Reliabilitas diterjemahkan dari kata *reliability*. Menurut Domino & Domino (2006, p.42) reliabilitas adalah hal yang dapat dipercaya. Kaplan & Saccuzzo (2005, p.21) menyatakan tentang reliabilitas, bahwa "*tests that are relatively free of measurement error are deemed to be reliable"*. Suryabrata (2005, p.28) berpendapat bahwa reliabilitas menunjukkan sejauhmana hasil pengukuran dengan alat tersebut dapat dipercaya. Hasil pengukuran harus reliabel dalam artian harus memiliki tingkat konsistensi dan kemantapan.

Cara yang sering digunakan adalah dengan rnenggunakan rumus reliabilitas alpha dari Cronbach yang merupakan salah satu upaya untuk rnenggabungkan rumus-rumus estimasi reliabilitas di bawah satu rumus yang umum. Cronbach sebagaimana dikutip Suryabrata (2005, p.37) mengusulkan koefisien Alpha yang rumusnya adalah sebagai berikut:

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum s_i^2}{s_X}\right)$$

Di mana α adalah koefisien reliabilitas, *k* adalah banyaknya butir tes, s^2_i varian dari skor butir tes dan s^2_X merupakan varian dari skor total. Berdasarkan runus di atas dapat dilihat bahwa semakin besar jumlah varian skor butir semakin tinggi juga estirnasi reliabilitas yang dihasilkan.

Menurut Nunnally & Bernstein (1994, p.245) untuk membuat keputusan individual, koefisien relibilitasnya minimal 0,6. Dalam pandangan Mardapi (2008, p.78) meskipun besaran indeks reliabilitas membentang dari 0 sampai 1, koefisien yang dapat diterima minimal 0,7. Koefisien reliabilitas berhubungan erat dengan kesalahan baku pengukuran. Hubungan antara koefisien reliabilitas dengan kesalahan pengukuran dinyatakan dengan persamaan berikut (Crocker & Algina, 1986, p.123).

$$\sigma_E = \sigma_X \sqrt{1 - r'_{XX}}$$

Simbol σE merupakan standar deviasi dari skor total. Hubungan antar indeks reliabilitas dengan kesalahan pengukuran berbanding terbalik. Semakin besar indeks reliabilitas, maka kesalahan pengukuran semakin kecil dan semakin kecil indeks reliabilitas maka kesalahan pengukuran semakin besar.

Teori Respon Butir

Pada banyak praktik pengukuran pendidikan, seseorang dapat menggunakan istilah-istilah deskriptif seperti kemampuan membaca dan kemampuan berhitung. Istilah-istilah tersebut merupakan variabel yang memiliki sifat laten atau sebagai suatu yang tidak teramati secara langsung. Variabel tersebut dapat dijelaskan dengan terlebih dahulu mendaftar atributnya, mengingat variabel ini tidak seperti ukuran tinggi atau berat yang dapat diukur secara langsung karena merupakan dimensi fisik. Atas dasar sifat laten seperti kemampuan atau abilitas (*ability*) inilah teori respons butir atau IRT terlahir (Baker, 2001, p.6).

Hambleton, Swarninathan, & Rogers (1991, p.7) menjelaskan bahwa IRT didasarkan pada dua konsep berikut. (a) performasi seorang subjek pada suatu butir dapat diprediksi atau dijelaskan oleh serangkaian faktor yang disebut trait, trait laten, atau kemampuan, (b) hubungan antara performasi subjek dalam suatu butir tes dan set trait yang mendasarinya dapat digambarkan dengan fungsi secara ajeg meningkat disebut fungsi karakteristik butir atau item *characterstic curve* (ICC).

Pendapat yang hampir sama disampaikan oleh Coley (2010, p.40) bahwa IRT didasarkan pada bentuk grafik yang berhubungan kemungkinan menjawab dengan benar untuk setiap item dengan kemampuan subjek. Pengembang tes yang mendasarkan pada IRT, mengasumsikan bahwa tanggapan terhadap item pada tes dapat dijelaskan oleh trait laten yang lebih sedikit jumlahnya dibandingkan dengan item tes. Memang, sebagian besar aplikasi dari teori mengasumsikan bahwa perkiraan trait laten tunggal untuk respon terhadap butir soal.

Dasar teori ini adalah model matematis tentang bagaimana peserta ujian pada tingkat kemampuan yang berbeda untuk menanggapi setiap karakteristik butir. Pengetahuan ini memungkinkan seseorang untuk membandingkan kinerja peserta ujian yang telah mengambil tes yang berbeda. Hal ini juga memungkinkan seseorang untuk menerapkan hasil analisis item untuk kelompok dengan tingkat kemampuan yang berbeda dibandingkan kelompok yang digunakan untuk analisis butir (Crocker & Algina, 2009, p.334).

Belbagai Asumsi IRT

Saifuddin Azwar (2010, pp.82-83) mengemukakan bahwa dalam IRT, probabilitas subjek untuk menjawab suatu item dengan benar tergantung pada kemampuan subjek dan karakteristik item. Secara tidak langsung dalam IRT juga memerlukan asumsi-asumsi pendukung, seperti unidimensionalitas dan independensi lokal.

Asumsi *unidimensionalitas* atau *unidimensionality* mensyaratkan bahwa setiap butir hanya mengukur satu macam ciri dikalangan peserta. Seperti diungkap (DeMars, 2010, p.38) bahwa unidimensionalitas berarti model memiliki trait tunggal untuk setiap pengujian, dan faktor-faktor lain yang mempengaruhi respons item diperlakukan sebagai kesalahan acak atau gangguan dimensi yang unik dan tidak dimiliki oleh item lainnya. Persyaratan butir unidimensi ini ditujukan untuk mempertahankan invariansi pada IRT. Invariansi berarti estimasi kemampuan yang diperoleh dari perangkat item yang berbeda akan senantiasa sama dan parameter item yang diperoleh dari kelompok subjek yang berbeda juga akan senantiasa sama, tentunya di luar kesalahan pengukuran. Butir tes yang mengukur lebih dari satu ciri peserta melalui lebih dari satu dimensi ukur dikenal dengan butir multidimensi.

Salah satu cara yang dapat digunakan untuk mengidentifikasi dimensi ukur adalah berdasarkan eigenvalue dari matriks korelasi antar-item dengan analisis faktor (DeMars, 2010, p.39; Lord, 1980, p.20). Melalui analisis faktor inilah dapat dipilih sejumlah butir ke dalam beberapa rumpun faktor.

Selain asumsi unidimensionalitas, dikenal juga asumsi independensi lokal atau local independence, yang menurut (DeMars, 2010, p.48), fokusnya adalah pada ketergantungan antar pasangan item. Artinya, apabila kemampuan-kemampuan yang mempengaruhi performasi dijadikan konstan, maka respons subjek pada pasangan item manapun juga akan independen secara statistik satu sama lain. Hal-hal yang dispesifikasikan dalam model yang merupakan faktor satu-satunya yang memiliki pengaruh terhadap respons yang ditunjukkan subyek.

Berbagai Model IRT

Pertama, IRT model logistik satu parameter. Model ini yang paling sering digunakan, mengingat prosedurnya yang sangat sederhana. Penyebutan satu parameter didasarkan pada karakteristik butirnya yang hanya ditunjukkan oleh statistik b_i, yang merupakan parameter tingkat kesukaran butir. Persamaan matematis untuk IRT model logistik satu parameter, seperti yang dipopulerkan oleh Hambleton, Swaminathan & Rogers (1991, p.12) adalah sebagai berikut.

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

Simbol $P_i(\theta)$ menunjukkan kemungkinan seorang subjek dengan tingkat kemampuan tertentu (θ baca *theta*) untuk menjawab butir *i* dengan benar. Simbol b_i menunjukkan parameter tingkat kesukaran butir *i*, simbol *n* menunjukkan banyaknya butir dalam tes, dan simbol *e* adalah eksponensial dengan angka konstan sebesar 2,718.

Paramater b_i adalah satu titik pada skala kemampuan di mana kemungkinan untuk menjawab benar sebesar 0,5. Semakin besar parameter b_i semakin besar pula kemampuan yang dituntut dan seorang subyek untuk memperoleh 50% peluang menjawab dengan benar. Ketika kemampuan sebuah kelompok ternyata memiliki rerata (0) dan deviasi standar (1), maka skor satandar *bi* akan berkisar dari -2,0 sampai + 2,0. Skor standar yang mendekati 2,0 menandakan bahwa butir soal mudah dan skor standar mendekati +2,0 berarti butir soal tersebut tergolong sulit untuk suatu kelompok subjek (Hambleton, Swaminathan & Rogers, 1991, p.13).

Kedua, IRT model logistik dua parameter dikembangkan berdasarkan pada distribusi normal kumulatif. Pengembangan model logistik dua parameter juga dimaksudkan untuk mengganti fungsi *ogive* normal dengan model yang lebih sederhana dan mudah untuk dianalisis. Secara matematis, DeMars, (2010, p.14) merumuskan IRT model logistik dua parameter sebagai berikut.

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

Model logistik dua parameter memiliki tambahan dibandingkan model dengan satu parameter. Model dua parameter terdapat factor *D* yang merupakan *factor* penskalaan dengan nilai yang konstan sebesar 1,7. Ternyata dalam model dua parameter perbedaan nilai $P_i(\theta)$ bagi kedua fungsi ogive normal dan fungsi logistik besarnya kurang dari 0,01 untuk semua nilai θ. Parameter a_i adalah indeks daya diskriminan butir atau indeks daya beda butir. Parameter ini proporsional terhadap slop kurva karakteristik butir di titik b_i pada skala kemampuan. Secara teoretis, parameter diskriminan ditetapkan pada rentang atau skala ($-\infty$, $+\infty$). Tetapi dalam prakteknya parameter negatif menghendaki butir tersebut tidak digunakan sedangkan parameter yang lebih besar dari 2 jarang terjadi. Dengan demikian, yang dilihat hanya parameter a_i yang besarnya antara 0 sampai 2.

IRT model logistik tiga parameter memiliki persamaan matematis yang diadaptasi dari yang dikemukakan Lord (1980, p.93), seperti berikut ini.

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

Satu parameter yang ditambahkan dalam model logistik tiga parameter adalah c_i yaitu parameter kemungkinan untuk menjawab benar secara kebetulan yang biasanya dikenal dengan *pseudo-chance level.* Dengan demikian dalam model logistik tiga parameter juga terdapat satu asumsi di mana seorang subyek yang memiliki kemampuan rendahpun bisa menjawab butir dengan benar. Hal ini biasanya berlaku untuk format tes pilihan ganda. Harga c_i biasanya diasumsikan akan lebih kecil dari

pada harga yang akan diperoleh bila subyek menjawab dengan tebakan secara acak.

Fungsi Informasi Butir dan Tes

Teori respons butir menyediakan metode yang ampuh untuk menjelaskan, memilih butir-butir tes yang baik dan membandingkan be-berapa macam tes. Metode tersebut tema-suk penggunaan fungsi informasi butir. Fungsi informasi butir adalah varians dari distribusi binomial yaitu hasil perkalian peluang menja-wab benar dengan peluang menjawab salah. Secara matematis, Hambleton & Swaminathan, (1985, p.105) merumuskan fungsi informasi butir sebagai berikut.

$$I(\theta, u_i) = \frac{(P_i^{'})^2}{P_i Q_i}$$

Simbol I (θ,u_i) menunjukkan fungsi informasi butir i pada tingkat kemampuan tertentu θ. *S*imbol P_i adalah peluang peserta dengan menjawab benar butir i dan $Q_i$adalah peluang peserta menjawab salah butir i. Fung-si informasi butir digunakan untuk menggam-barkan kekuatan butir tes serta untuk memban-dingkan beberapa butir tes pada indikator yang serupa.

Ada fungsi informasi butir, tentu saja ada fungsi informasi tes yang pada dasarnya dirumuskan sebagai jumlah dari fungsi infor-masi tes. Oleh karenanya, tinggi rendahnya fungsi informasi tes ditentukan oleh fungsi informasi butir. Fungsi informasi tes secara matematis dapat dituliskan (Hambleton & Swaminathan, 1985, p.104) sebagai berikut.

$$I_i(\theta) = \sum_{i=1}^{n} \frac{\left[P_i^{'}(\theta)\right]^2}{P_i(\theta) Q_i(\theta)},\ i = 1, 2, 3, \ldots n$$

Nilai-nilai yang dihasilkan oleh fungsi informasi butir dan tes pada dasarnya adalah nilai estimasi, sehingga kebenaran yang dihasil-kannya memiliki probabilitas. Kesalahan peng-ukuran memiliki peran yang penting dalam penentuan keakuratan hasil estimasi. Di dalam IRT, kesalahan pengukuran sangat berhubungan dengan fungsi informasi tes, yang secara matematis kesalahan pengukuran dirumuskan sebagai berikut (Hambleton & Swaminathan, 1985: 104):

$$SE_i(\theta) = \frac{1}{\sqrt{I_i(\theta)}},\ i = 1, 2, 3, \ldots n$$

SEi(θ) adalah kesalahan pengukuran pada butir i dan tingkat kemampuan subjek θ. Semakin besar fungsi informasi, maka kesalah-an pengukuran semakin kecil, demikian juga sebaliknya.

METODE PENELITIAN

Jenis Penelitian

Penelitian ini termasuk jenis penelitian survei, yang mengambarkan karakteristik butir soal UN bahasa Inggris SMK TA 2010/2011 di Kabupaten Tabalong. Penelitian ini bersifat *ex-post facto*, karena peneliti tidak melakukan perlakuan apapun terhadap variabel penelitian.

Waktu dan Tempat

Waktu dilaksanakannya penelitian adalah pada bulan Mei-Agustus 2010, diawali dengan membuat proposal. Tempat dilaksanakannya penelitian adalah Kabupaten Tabalong, Provinsi Kalimentan Selatan. Dokumen mengenai penyelenggaraan dan hasil UN Bahasa Inggris SMK diambil dari dinas pendidikan setempat.

Subjek dan Objek

Jumlah paket soal UN Bahasa Inggris SMK yang digunakan di kabupaten Tabalong berjumlah 5 buah paket soal, namun yang dianalisis hanya paket 15 yang telah mewakili paket 27, paket 34, paket 41, dan paket 59. Hal ini karena pada dasarnya model soal sama hanya pada paket yang lain diacak nomor soalnya. Semua paket soal digunakan pada SMK/MK berstatus negeri dan swasta di Kabupaten Tabalong dengan jumlah siswa yang menggunakan sebanyak 784 orang siswa.

Teknik Pengumpulan Data

Penelitian ini menggunakan data sekun-der, sehingga pengumpulan data dalam bentuk dokumen lembar jawaban atau respons siswa terhadap soal-soal yang terdapat pada perangkat tes. Akan tetapi, dikarenakan lembar-lembar jawaban tersebut telah berada di Jakarta, maka respons siswa terhadap soal didapatkan dari data yang dimiliki Dinas Pendidikan dan Kebuda-yaan Provinsi Kalimantan Selatan pada bagian pendidikan menengah.

Prosedur Penelitian

Sesuai dengan tujuan penelitian yang bermaksud mengungkap karakteristik butir soal UN berdasarkan analisis butir soal secara kuali-tatif dan kuantitatif, serta memetakan kualitas pembelajaran bahasa Inggris pada level kabupa-

ten, maka terdapat tiga tahap setelah menyusun proposal. Pertama, mengumpulkan dokumen naskah soal UN bahasa Inggris SMK TA 2010/ 2011 yang digunakan di Kabupaten Tabalong beserta hasil pemindaian LJK dan daftar kolektif hasil UN. Kedua, menghadirkan 12 orang guru bahasa Inggris guna melakukan analisis butir soal secara kualitatif (aspek materi, konstruksi, dan bahasa/budaya) terhadap butir soal pada naskah soal UN bahasa Inggris. Ketiga, melakukan analisis butir soal secara kuntitatif memanfaatkan komputer paket program ITEMAN 3.00 untuk CTT dan EIRT 1.0.9 untuk IRT. Memanfaatkan juga paket program MS Excel dan SPSS 16.0 untuk memeta kualitas pembelajaran bahasa Inggris.

Teknik Analisis Data

Analisis kuantitatif memanfaatkan paket program ITEMAN 3.00 untuk pendekatan CTT yang hasil analisisnya memuat karakteritik butir soal berupa (1) tingkat kesukaran, (2) daya beda, (3) efektifitas distraktor. Sedangkan statistik perangkat tes yang dihasilkan dari program ini adalah mean, median, indeks kehandalan, kemencengan dan kesalahan baku pengukuran. Pendekatan IRT memanfaatkan paket program EIRT 1.0.9 pendekatan yang dilakukan dengan 1, 2, maupun 3 parameter logistik. Program ini akan menghasilkan lima macam output, yaitu: *measure* butir soal dan *measure* peserta ujian, *test of fit*, *test of independence local*, *parameter estimate*, *estimate latent variable*, *total and item characteristic curve*, dan *total and item information function*. Paket program tersbut dipilih karena mudah cara mengoperasikannya dan efisien (Germain, Valois & Abdous, 2007, pp.1-27). Paket program SPSS 16.0 juga dimanfaatkan untuk memeta kualitas pembelajaran bahasa Inggris dengan analisis varian dilanjutkan *Duncan Multiple Ranks Test* (DMRT).

HASIL PENELITIAN

Analisis Kualitatif

Analisis kualitatif terhadap butir soal ini disebut juga telaah butir soal dan butir soal yang dimaksud adalah konten butir soal UN SMK TA 2010/2011 untuk mata pelajaran bahasa Inggris yang digunakan di Kabupaten Tabalong Kalimantan Selatan. Objek telaah adalah aspek materi, konstruksi dan bahasa/budaya yang terdapat dalam setiap butir soal naskah UN. Subjek atau penelaah adalah guru bahasa Inggris di wilayah kabupaten tersebut berjumlah 12 orang guru.

Hasil telaah aspek materi butir soal menunjukkan bahwa secara umum butir soal UN dapat dinyatakan telah sesuai indikator, pilihan jawaban homogen dan logis, dan hanya ada satu kunci jawaban. Namun, dari 50 butir soal ada 3 (6%) butir, yaitu butir soal nomor 43, 44, dan 45 yang kurang sesuai dengan kompetensi. Ketiga soal tersebut mestinya memuat bacaan bukan tabel waktu, karena kompetensi yang diukur adalah membaca. Aspek konstruksi butir soal UN telah dirumuskan secara singkat, jelas, tegas, dan jawaban tidak bergantung pada butir soal yang lain. Kelemahannya pada butir 1, 2, 3, dan 4, berarti dari 50 butir soal ada 4 (8%) butir yang menunjukkan gambar, grafik, tabel, diagram atau sejenisnya kurang jelas dan belum berfungsi dengan baik. Pada aspek bahasa/budaya, hasil telaah menunjukkan bahwa butir soal UN telah sesuai kaidah bahasa Inggris yang benar, komunikatif, tidak tabu, dan tidak mengulang kata yang tidak perlu.

Analisis Kuantitatif

Analisis kuantitatif terhadap butir soal UN SMK TA 2010/2011 untuk mata pelajaran bahasa Inggris yang digunakan di Kabupaten Tabalong Kalimantan Selatan ini menerapkan CTT dan IRT. Objek telaah berdasarkan pendekatan CTT adalah karakteritik butir berupa tingkat kesukaran, daya pembeda, efektivitas distraktor, dan estimasi reliabilitas, sedang pendekatan IRT adalah kecocokan model, fungsi informasi, dan kesalahan baku pengukuran. Subjeknya adalah peserta UN yang berjumlah 784 yang sumber datanya diperoleh dari dokumen hasil pemindaian lembar jawaban komputer dan dokumen daftar kolektif hasil UN.

Tabel 2. Karakteristik Butir Tes Pendekatan CTT

No butir	p_i	r_{PBis}	CK	DF	No butir	p_i	r_{PBis}	CK	DF
1	0,955	0,196	-	-	26	0,404	0,408	-	3
2	0,649	-0,073	√	3	27	0,511	0,484	-	2
3	0,781	0,250	-	1	28	0,204	0,572	-	3
4	0,395	0,050	√	2	29	0,376	0,163	-	2
5	0,813	0,240	-	2	30	0,705	0,010	√	1
6	0,870	0,202	-	1	31	0,526	0,289	-	2
7	0.756	-0.056	√	1	32	0,588	0,077	-	3
8	0,892	0,344	-	1	33	0,867	0,234	-	-
9	0,323	0,323	-	3	34	0,571	0,373	-	3
10	0,527	0,334	-	2	35	0,601	0,159	-	3
11	0,653	0,176	-	2	36	0,353	0,405	-	3
12	0,923	0,116	-	-	37	0,457	0,418	-	3
13	0,532	0,235	-	3	38	0,497	0,122	-	3
14	0,640	0,241	-	2	39	0,181	0,015	√	2
15	0,619	0,526	-	1	40	0,509	0,276	-	2
16	0.596	0,625	-	3	41	0,348	0,547	-	3
17	0,432	0.411	-	2	42	0,375	0,442	-	3
18	0,631	0,527	-	2	43	0,446	0,061	-	3
19	0,222	-0,090	√	3	44	0,540	-0,022	√	3
20	0,380	0,450	-	3	45	0,895	0,049	-	-
21	0,829	0,298	-	2	46	0,807	0,349	-	2
22	0,750	0,089	-	2	47	0,718	0,227	-	3
23	0,381	0,547	-	1	48	0,917	0,167	-	-
24	0,668	0,073	-	1	49	0,503	0,436	-	3
25	0,708	0,513	-	1	50	0,372	0,243	√	2

Keterangan: p_i = indeks kesukaran, r_{PBis} = indeks daya pembeda,
CK = perlu cek kunci jawaban (√), dan DF = jumlah distraktor berfungsi

Hasil analisis pendekatan CTT untuk indeks kesukaran menunjukkan bahwa dari 50 butir soal, ada 31 (62%) butir berkategori sedang, 16 (32%) butir berkategori mudah, dan 3 (6%) butir berkategori sukar. Dilihat dari indeks daya pembeda, ada 17 (34%) butir berkatogori cukup, 15 (30%) butir berkategori buruk, 14 (28%) butir berkategori baik, 4 (8%) berkategori sangat buruk, dan tidak ada yang berkategori sangat baik. Ditinjau dari efektivitas distraktor serta kunci jawaban, ada 36 (72%) butir distraktornya berfungsi efektif, dan 14 (28%) butir distraktornya tidak berfungsi efektif, serta ada 8 (16%) butir yang memerlukan cek kunci jawaban. Estimasi reliabilitas untuk 50 butir soal UN diperolah angka koefisien sebesar 0,729 yang berarti reliabilitasnya tinggi. Setelah dilakukan seleksi butir soal berdasarkan indeks kesukaran, indeks daya pembeda, dan efektivitas distraktor serta kualitas kunci jawaban diperoleh ada 30 (60%) butir yang dapat dinyatakan ”kurang baik” dan ada 20 (40%) butir yang dapat dinyatakan ”baik”. Apabila butir soal yang dinyatakan kurang baik diseleksi atau dihilangkan, koefisien reliabilitasnya menjadi 0,784 yang berarti estimasi reliabilitasnya mengalami peningkatan meskipun masih termasuk tinggi.

Tabel 3. Karakteristik Butir Tes Pendekatan IRT

No	1PL	2PL		3PL			No	1PL	2PL		3PL		
Butir	b_i	a_i	b_i	a_i	b_i	c_i	Butir	b_i	a_i	b_i	a_i	b_i	c_i
1	-3.15	0.61	-3.29	0.72	-3.38	0.17	26	0.50	0.53	0.55	3.27	0.39	0.26
2	-0.60	0.10	-3.41	0.07	-1.22	0.27	27	0.02	1.20	0.02	0.93	-0.60	0.06
3	-1.30	0.25	-3.03	0.31	-2.66	0.18	28	1.56	1.82	0.89	1.00	0.00	0.20
4	0.54	0.06	3.67	0.09	4.11	0.12	29	0.63	0.11	2.69	1.00	0.00	0.20
5	-1.51	0.35	-2.56	0.40	-2.65	0.15	30	-0.89	0.10	-5.07	0.04	-1.10	0.39
6	-1.97	0.24	-4.60	0.30	-3.98	0.20	31	-0.04	0.35	-0.13	0.30	-0.21	0.15
7	-1.16	0.21	-3.14	0.18	-3.13	0.22	32	-0.32	0.05	-3.43	0.06	-0.15	0.20
8	-2.18	0.87	-1.78	0.98	-2.22	0.13	33	-1.96	0.27	-4.13	0.38	-3.32	0.20
9	0.88	0.19	2.33	0.30	1.65	0.10	34	-0.25	0.57	-0.29	0.61	-0.87	0.06
10	-0.05	0.18	-0.30	0.29	-0.21	0.14	35	-0.37	0.08	-2.69	0.08	-0.48	0.20
11	-0.62	0.10	-3.36	0.08	-1.47	0.25	36	0.73	0.64	0.71	3.23	0.44	0.22
12	-2.57	0.27	-5.51	0.28	-5.32	0.21	37	0.27	1.08	0.20	1.32	-0.25	0.15
13	-0.07	0.13	-0.54	0.18	0.15	0.17	38	0.09	0.07	0.21	1.00	0.00	0.20
14	-0.56	0.13	-2.49	0.17	-1.41	0.19	39	1.70	0.19	4.71	1.48	2.05	0.18
15	-0.46	0.76	-0.42	1.04	-0.94	0.08	40	0.02	0.57	0.00	0.73	-0.27	0.18
16	-0.36	1.43	-0.21	1.93	-0.85	0.04	41	0.75	1.16	0.52	1.00	0.00	0.20
17	0.37	0.68	0.34	1.00	0.00	0.20	42	0.63	0.83	0.52	1.41	-0.08	0.11
18	-0.52	1.12	-0.36	1.05	-1.04	0.04	43	0.31	0.06	1.86	0.10	2.35	0.12
19	1.44	0.11	6.26	1.00	0.00	0.20	44	-0.08	0.04	-1.70	0.03	1.00	0.13
20	0.60	0.76	0.52	1.64	0.14	0.17	45	-2.27	0.26	-5.02	0.28	-4.74	0.22
21	-1.64	0.34	-2.80	0.49	-2.32	0.23	46	-1.49	0.81	-1.28	0.79	-1.85	0.12
22	-1.12	0.11	-5.38	0.11	-3.30	0.32	47	-0.93	0.29	-1.87	1.45	-0.28	0.54
23	0.61	2.14	0.34	1.00	0.00	0.20	48	-2.52	0.34	-4.27	0.43	-3.83	0.20
24	-0.69	0.13	-2.97	0.26	-1.25	0.24	49	0.07	1.13	0.06	1.22	-0.59	0.04
25	-0.90	1.68	-0.50	1.69	-1.18	0.05	50	0.64	0.21	1.53	0.27	1.14	0.09

Keterangan: a_i = indeks daya pembeda c_i = *guessing* b_i = indeks kesukaran

Analisis pendekatan IRT untuk kecocokan butir soal, dari 50 butir soal UN yang cocok dengan model 3PL sebesar 27 (54%) butir, model 2PL sebesar 24 (48%) butir, dan model 1PL sebesar 21 (42%) butir. Ditinjau dari fungsi informasinya, model 3PL lebih tinggi dibandingkan model 2PL dan 1PL. Sedangkan apabila ditinjau dari kesalahan baku pengukurannya, model 3PL lebih rendah dibandingkan 2PL dan 1PL. Berdasarkan kecocokan model, fungsi informasi, dan kesalahan baku pengukuran, maka model 3PL dipandang lebih presisi digunakan dalam mengestimasi abilitas peserta UN sesuai karakteritik butir soal UN, meskipun sampel data kurang memenuhi persayaratan. Memilih IRT model 3PL berarti mendasarkan karakteristik butir pada tingkat kesukaran, daya pembeda, dan menjawab benar secara kebetulan (*guessing*).

PEMBAHASAN

Perbandingan Analisis Kualitatif dan Kuantitatif

Hasil analisis kualitatif menjunjukkan bahwa secara umum butir soal UN memiliki butir yang karakteristiknya dapat dinyatakan baik berdasarkan aspek materi, konstruksi, dan bahasa/budaya. Meski sedikit terdapat kelemahan pada aspek materi, yaitu butir nomor 43, 44, dan 45 yang dinyatakan belum sesuai kompetensi yang diukur. Sedikit kelemahan juga terjadi pada aspek konstruksi, di mana butir nomor 1, 2, 3, dan 4 memuat gambar, grafik, tabel, diagram atau sejenisnya kurang jelas dan

belum berfungsi dengan baik. Hal ini menunjukkan bahwa butir soal kurang baik berdasarkan analisis kualitatif ada 7 (14%) butir, sedangkan 43 (86%) butir yang lainnya dapat dianggap baik.

Hasil analisis kuantitatif pendekatan CTT menunjukkan ada 20 (40%) butir yang dapat dinyatakan baik karakteritiknya, sementara pendekatan IRT model 3PL menunjukkan ada 27 (54%) butir yang dapat dinyatakan baik karakteritiknya. Pada analisis kuantitatif sendiri ternyata masing-masing pendekatan menunjukkan perbedaan hasil sehingga memerlukan pembahasan tersendiri agar lebih jelas.

Analisis kualitatif bersumber pada telaah butir soal secara teoretis atau konten butir soal pada naskah soal UN, sedangkan analisis kuantitatif bersumber dari respons siswa atas butir soal UN yang ada. Ketujuh butir yang dinyatakan kurang baik karakteritiknya berdasarkan hasil analisis kualitatif juga tidak baik berdasarkan hasil analisis kuantitatif pendekatan CTT, namun tidak demikian untuk pendekatan IRT. Pada pendekatan IRT, butir nomor 1, 2, dan 45 dapat dianggap baik karakteritiknya karena cocok dengan model 3PL, ini artinya ketiga butir itu memiliki tingkat kesukaran, daya pembeda, dan *guessing* yang dianggap mampu mengestimasi abilitas yang sesungguhnya dari peserta UN.

Perbandingan CTT dan IRT

Agar mengetahui perbandingan antara hasil analisis pendekatan CTT dan IRT, maka digunakan koefisien determinasi menggunakan analisis korelasi dan regresi. Pada tingkat kesukaran butir soal, koefisien determinasi antara hasil analisis CTT dan IRT adalah 0.661, artinya ada kesesuaian sekitar 66,1% dalam hal tingkat kesukaran. Pada daya pembeda butir soal, koefisien determinasi antara hasil analisis CTT dan IRT adalah 0,237, artinya ada kesesuaian sekitar 23,7% dalam hal daya pembeda. Hasil analisis IRT model 3PL terdapat parameter *guessing* sebagai salah satu yang menggambarkan karakteritik butir, sedangkan CTT tidak secara spesifik menjelaskan karakteritik ini. CTT memiliki kelemahan bahwa karakteritik butir soal sangat tergantung pada peserta yang mengerjakan soal itu, dan estimasi abilitas peserta juga sangat tergantung pada karakteritik butir soal yang dikerjakannya. IRT memiliki keunggulan dapat menutupi kelemahan yang ada pada CTT tersebut, karena didasarkan pada konsep invarians.

Mempertimbangkan keunggulan yang dimiliki IRT dibandingkan CTT, maka dalam melakukan pemetaan kualitas pembelajaran bahasa Inggris berbasis hasil UN didasarkan atas estimasi abilitas peserta untuk butir-butir soal yang dianggap baik atas hasil analisis pendekatan IRT.

Peta Kualitas Pembelajaran Bahasa Inggris Berbasis Hasil UN

Hasilnya menunjukkan SMKN 1 Haruai, SMKN 1 Muara Uya, dan SMKN 1 Tanjung memiliki kualitas yang dapat dinyatakan seragam dengan lainnya. Empat sekolah yang juga dapat dinyatakan seragam kualitas pembelajaran bahasa Inggrisnya adalah SMK An Noor Paliat, SMKN 1 Haruai, SMKN 1 Muara Uya, dan SMKN 1 Tanjung, meski demikian SMK Tabalong dapat dianggap lebih unggul diabandingkan SMK An-Noor Paliat dan SMKN 1 Banua Lawas. Secara berturut-turut dari tinggi ke rendah terpeta kualitas pembelajaran bahasa Inggris di Kabupaten Tabalong adalah SMK Tabalong paling unggul, disusul SMKN 1 Tanjung, SMKN 1 Muara Uya, SMKN 1 Haruai, SMKN 1 Banua Lawas, dan SMK An Noor Paliat.

SIMPULAN DAN SARAN

Simpulan

Karakteristik soal UN bahasa Inggris SMK di Kabupaten Tabalong berdasarkan pendekatan kualitatif dan kuantitatif, serta hasil pemetaan kualitas pendidikan adalah sebagai berikut.

Pertama, pendekatan kualitatif yang terdiri atas tinjauan aspek substansi, konstruksi, dan bahasa termasuk dalam kategori baik. Soal UN telah memenuhi hampir semua kriteria dari aspek materi, konstruksi, dan bahasa. Kelemahannya pada penggunaan gambar, grafik, tabel, diagram atau sejenisnya yang belum berfungsi dengan baik, serta dari aspek materi ditemukan beberapa soal tidak sesuai dengan kompetensi yang diukur.

Kedua, karakteristik soal UN berdasarkan analisis kuantitatif dengan teori tes klasik atau ditinjau dari indeks kesukaran, indeks daya pembeda, keefektifan ditraktor hasilnya hanya 40% atau 20 butir yang dapat dikategorikan baik dan dari estimasi reliabilitas termasuk kategori tinggi. Karakteristik soal UN ditinjau dari model IRT, khusus model 3PL lebih relevan dibandingkan 1PL maupun 2PL dengan tingkat kecocokan 54%. Penggunaan model IRT 3PL dipilih

mengingat informasi yang dipeoleh lebih banyak dan kesalahan pengukuran lebih kecil dari model yang lain. Ini artinya dalam soal UN tersebut telah mempertimbangkan tingkat kesukaran, daya pembeda, dan faktor menebak.

Ketiga, pemetaan kualitas pendidikan di kabupaten Tabalong berdasarkan hasil UN Bahasa Inggris TA 2010/2011, menunjukkan semua SMK dalam lingkup Kabupaten Tabalong di bawah rata-rata kabupaten. SMK Tabalong paling unggul di antara SMK yang lainnya, disusul SMKN 1 Tanjung, SMKN 1 Muara Uya, SMKN 1 Haruai, SMKN 1 Banua Lawas, dan SMK An Noor Paliat.

Saran

Pertama, rumusan pengembangan kebijakan penyusunan butir-butir soal UN di Kabupaten Tabalong disarankan untuk disesuaikan dengan situasi dan kondisi wilayah setempat.

Kedua, guru-guru mata pelajaran bahasa Inggris di kabupaten Tabalong atau yang mewakilinya disarankan terlibat langsung dalam pembuatan soal sebagai bagain dari pengembangan kemampuan guru. Hal ini lebih menjamin kualitas isi dari perangkat tes. Selain itu, setiap daerah memiliki hambatan-hambatan masing-masing dalam proses pembelajaran serta untuk menumbuhkan rasa tanggungjawab dari para guru dalam rangka melaksanakan UN SMK yang lebih baik dan jujur.

Ketiga, pemerintah hendaknya memperhatikan daerah-daerah yang masih tertinggal dalam bidang pendidikan dan teknologi, khususnya penguasaan bahasa Inggris, diberikan bantuan sebagaimana hasil pemerataan kualitas pendidikan dari hasil UN bahasa Inggris SMK.

DAFTAR PUSTAKA

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

Anthony, L. (2012). *English specific purposes: What does it mean? Why is it different?* Artikel [*Version electronic*]. Diambil pada tanggal 8 Januari 2013. http://www.antleb.sci.waseda.ac.jp/abstract/ESParticels.html

Azwar, Saifuddin. (2012). *Penyusunan skala psikologi*. Yogyakarta: Pustaka Pelajar.

_____, (2010). *Dasar-dasar psikometri*. Yogyakarta: Pustaka Pelajar.

Baker, F (2001). *The basics of item response theory*. Boston: Lawrence Rudner

Basturkman, H. (2006). *Idea and option in english for specific purposes.* New Jersey: Laurence Erlbaum Associates Publishers.

Chadha, N. K., (2009). *Applied psychometry*. New Delhi: SAGE Publication Inc.

Coaley, K. (2010). *An introduction to psychological assessment and psychometrics*. London: SAGE Publication Inc.

Crocker, L, & Algina, J. (2008). *Introduction to classical and modern test theory.* Ohio: Cengage Learning.

Dali S. Naga, (1992). *Pengantar teori skor pada pengukuran pendidikan*. Jakarta: Gunadharma.

DeMars, C, (2010). *Item response theory*. New York: Ox ford University Press, Inc.

Depdiknas, (2009). *Ujian nasional*: Saat ini dan masa mendatang. Pusat Penilaian Pendidikan, Badan Penelitian dan Pengembangan, Departemen Pendidikan Nasional RI.

_____. *(2006a). Peraturan Menteri Pendidikan Nasional Nomor 23 Tahun 2006 tentang Standar Kompetensi Lulusan*

_____. *(200b). Peraturan Menteri Pendidikan Nasional Nomor 22 Tahun 2006 tentang Standar Isi*

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

Ebel, R.L, & Frisbie, D.A, (1991). *Essential of educational measurement*. New Delhi: Prentice Hall

Germain, S, Valois P, & Abdous, B. (2007). *Manual EIRT 1.0.9: Item response theory assistant for excel*. Artikel [*Version electronic*]. Diambil pada tanggal 7 Januari 2013. http://libirt.sourceforge.net/ eirt-en/index.html

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

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. London: SAGE Publication Inc.

Hutchinson, T., and Water, A. (1991). *English for specific purposes: A learning-centred approach*. New York. Cambridge University Press.

Kaplan, R.M, & Saccuzzo, D.P. (2005). *Psychological testing: Principles, applications, and issues*. Canada: Wadsworth Thomson Learning Inc

Mardapi, Djemari. (2008). *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta: Mitra Cendekia.

Nunnally, J.C, & Berensten, I.H. (1994). *Psychometric theory*. New York: McGrow Hill Inc.

Premono, Agung. (2010). Kompetensi keahlian sekolah menengah kejuruan: Antara kebijakan dan realita. *Jurnal Penddikan Penabur, 15, 9, pp.50–61*.

Sieroeka, H. (2008). *The role of the ESP teacher*. Business English [*Version electronic*]. Diambil pada tanggal 9 Januari 2013. http://www.jezykangielski. org/theroleoftheespteacher.pdf

Suryabrata, Sumadi. (2005). *Pengembangan alat ukur psikologis*. Yogyakarta: Andi Offset.

Tarigan, HG. (2008). *Menyimak sebagai suatu keterampilan berbahasa*. Bandung Angkasa.

Zaman, et al. (2010). Analysis of multiple choice items and the effect of items sequencing on difficulty level in the test of mathematics. *European Journal of Social Sciences, 17/1, 62–66.*

Zulaiha, Rahmah. (2008). *Analisis soal secara manual*. Pusat Penilaian Pendidikan, Badan Penelitian dan Pengembangan, Departemen Pendidikan Nasional RI.