Implementasi Teori Responsi Butir (*Item Response Theory*) pada Penilaian Hasil Belajar Akhir di Sekolah

Sudaryono
sudaryono2@yahoo.com

ABSTRAK: Pengukuran pendidikan meliputi pengukuran hasil belajar dari berbagai bidang, tergantung objek hasil belajar apa yang ingin diukur. Oleh karena itu, yang menjadi permasalahan dalam artikel ini: 1) apakah teori responsi butir atau teori tes modern bisa menutupi kelemahan-kelemahan yang ada pada teori tes klasik; 2) bagaimana implementasi teori responsi butir dalam mengatasi permasalahan-permasalahan ujian nasional sehingga tidak ada kelompok yang diuntungkan dan kelompok yang dirugikan akibat pengukuran yang tidak adil? Tujuan dari penulisan artikel ini adalah menjelaskan implementasi teori responsi butir dalam menutupi kelemahan yang ada pada teori tes klasik dan mengatasi permasalahan ujian nasional, sehingga tidak ada kelompok yang dirugikan maupun diuntungkan akibat pengukuran yang tidak adil. Teori responsi butir merupakan alternatif pilihan yang bertujuan melepaskan diri dari ketergantungan tes yang diberikan dengan sampel peserta tes. Dalam hal ini walaupun soal-soal tersebut dikerjakan oleh siswa yang pandai atau siswa yang kurang pandai, indikasi tingkat kesukaran suatu soal tetap tidak berubah. Ada tiga asumsi yang harus dipenuhi dalam teori response butir, yaitu: 1) unidimensi; 2) independensi lokal; dan 3) invariansi sedangkan karakteristik butir ada tiga, yaitu: 1) taraf sukar butir; 2) daya beda butir; dan 3) tingkat kebetulan betul pada butir. Untuk mengukur kemampuan peserta tes yang sangat beragam di Indonesia, seperti Ujian Nasional, seharusnya digunakan juga ujian atau tes yang berbeda tingkat kesukaran soalnya, supaya adil dan juga akurat hasilnya. Peserta tes atau ujian yang mengerjakan tes atau ujian yang berbeda tingkat kesukaran soalnya, tetap bisa dibandingkan kemampuannya, asalkan soal-soal dalam ujian tersebut berasal atau diambil dari bank soal yang sudah dikalibrasi dengan konsep *item response theory*.

Kata Kunci: *teori responsi butir, unidimensi, bank soal, independensi lokal, invariansi, taraf sukar butir, tingkat kesukaran soal.*

ABSTRACT: Educational measurement, consisted measurement of learning outcomes from a variety of fields, depending on the object of learning what to measure. Therefore, the problem raised in this paper are: 1) whether the item response theory or theories of modern tests can cover weaknesses that exist in classical test theory, 2) how the item response theory implementations in addressing issues of national exams so that no advantaged groups and disadvantaged groups as a result of measurement that is not fair? The purpose of writing this article is to explain the implementation of item response theory in a cover up weaknesses in classical test theory and address the issues of national examinations, so that no group is disadvantaged or advantaged as a result of measurement that is not fair. Item response theory is an alternative option that aims to break away from dependence on a given test with a sample of test participants. In this case, although the questions are done by a brilliant student or students who are less intelligent, an indication of the level of difficulty of a problem remains unchanged. There are three assumptions that must be met in item response theory, namely: 1) unidimention; 2) local independence, and 3) invariance. While there are three characteristic points, namely: 1) the item difficulty, 2) the different grains, and 3) the level of true coincidence in point. To measure the ability of the test participants are very diverse in the premises, such as the National Examination, should be used is also an examination or test different levels of difficulty because, to be fair and accurate results. Participants test or exam is working on a test or exam because of different levels of difficulty, it can be compared to his ability, provided the questions in the exam are derived or extracted from a question bank that has been calibrated with the concept of item response theory.

Keywords: *item response theory, unidimention, local independence, invariance, item difficulty, item bank, the dificulty level of items.*

Pendahuluan

Ujian Nasional merupakan salah satu penilaian eksternal yang digunakan pemerintah untuk mengumpulkan data pencapaian prestasi belajar peserta didik, sejauh mana prestasi belajar peserta didik mencapai Standar Kompetensi Lulusan (SKL). Di sekolah peserta didik seharusnya sudah terbiasa dengan penilaian hasil belajar yang dilakukan oleh guru sekolah. Sebagaimana diamanatkan oleh Peraturan Pemerintah Nomor 19 Tahun 2005 tentang Standar Nasional Pendidikan Pasal 63 ayat (1): Penilaian pendidikan pada jenjang pendidikan dasar dan menengah terdiri atas: 1) penilaian hasil belajar oleh pendidik; 2) penilaian hasil belajar oleh satuan pendidikan; dan 3) penilaian hasil belajar oleh pemerintah (Wibowo, 2011).

Penilaian hasil belajar oleh pendidik dilakukan secara berkesinambungan untuk memantau proses, kemajuan, dan perbaikan hasil dalam bentuk ulangan harian, ujian tengah semester, ujian akhir semester, dan ujian kenaikan kelas. Penilaian hasil belajar oleh pendidik digunakan untuk menilai pencapaian kompetensi peserta didik; bahan penyusunan laporan hasil belajar; dan memperbaiki proses pembelajaran. Penilaian hasil belajar oleh satuan pendidikan bertujuan menilai pencapaian standar kompetensi lulusan untuk semua mata pelajaran. Penilaian hasil belajar oleh pemerintah dalam bentuk ujian nasional bertujuan untuk menilai pencapaian kompetensi lulusan secara nasional pada mata pelajaran tertentu dalam kelompok mata pelajaran ilmu pengetahuan dan teknologi. Ujian nasional dilakukan secara objektif, berkeadilan, dan akuntabel.

Hasil ujian nasional digunakan sebagai salah satu pertimbangan untuk: 1) pemetaan mutu program dan/atau satuan pendidikan; 2) dasar seleksi masuk jenjang pendidikan berikutnya; 3) penentuan kelulusan peserta didik dari program dan/atau satuan pendidikan; dan 4) pembinaan dan pemberian bantuan kepada satuan pendidikan dalam upayanya untuk meningkatkan mutu pendidikan.

Dalam kaitan ini, persoalan yang akan disoroti dan dikaji adalah dari aspek penggunaan tes yang dirancang sedemikian rupa sehingga menimbulkan pertanyaan, sejauh mana tes tersebut telah sesuai dengan kemampuan siswa yang menjawabnya? Hal ini berhubungan dengan tingkat kevalidan atau kesahihan tes yakni sejauh mana tes tersebut benar-benar mengukur aspek yang diukur. Aiken (1988: 103) mendefinisikan validitas sebagai berikut *Validity of a test has been defined as the extent to wich the test measures what it was designed to measures*. Dalam penyusunan tes yang dirancang sebagai tes standar untuk mengungkapkan kemampuan peserta tes, maka analisis validitas dan reliabilitas butir sangat penting dilakukan. Bagi yang memerlukan informasi mengenai validitas dan reliabilitas item dalam mengestimasi validitas dan reliabilitas perangkat item yang bakal terpilih sebagai tes, dapat menggunakan fungsi indeks reliabilitas dan indeks validitas item yang bertujuan untuk meningkatkan reliabilitas dan validitas tes secara keseluruhan (Azwar, 2001). Dalam kaitan ini, tinjauan diarahkan pada pengkajian penerapan tes modern yakni teori responsi butir (*item response theory*) dalam penilaian hasil belajar peserta didik dengan segala atribut dan persyaratan-persyaratan yang dimilikinya.

Pada prinsipnya, pengukuran bertujuan untuk mengetahui karakteristik suatu objek yang akan diukur. Khususnya, pengukuran pendidikan meliputi pengukuran hasil belajar mencakup bermacam bidang, tergantung objek hasil belajar apa yang ingin diukur. Permasalahan dalam tulisan ini adalah: 1) apakah teori responsi butir atau teori tes modern bisa menutupi kelemahan-kelemahan yang ada pada teori tes klasik; 2) bagaimana implementasi teori responsi butir dalam mengatasi permasalahan-permasalahan ujian nasional sehingga tidak ada kelompok yang diuntungkan dan kelompok yang dirugikan akibat pengukuran yang tidak adil? Sedangkan yang menjadi tujuan penulisan artikel ini adalah: 1) untuk memberikan kajian secara singkat implementasi item responsi teori dalam pengembangan butir soal ujian nasional sehingga dapat berlaku adil untuk semua peserta didik; 2) memberikan masukan bagi sekolah dalam membuat butir soal yang sesuai dengan kaidah-kaidah pengukuran modern dengan menggunakan teori responsi butir.

Kajian Literatur dan Pembahasan

Penskoran Klasik dan Modern

Berdasarkan taksonomi psikologi belajar, maka karakteristik objek berkaitan dengan aspek kognitif, afektif dan psikomotorik. Secara khusus, pengukuran aspek kognitif diukur melalui uji tes, sedangkan pengukuran aspek afektif diukur dengan kuesioner, angket, wawancara, atau melalui pengamatan, sementara aspek psikomotorik diukur dengan

pengamatan langsung melalui praktik terhadap sesuatu keterampilan (*skill*) khusus dari peserta didik. Objek yang diukur dalam pendidikan antara lain: siswa, mahasiswa, guru/dosen. Untuk mendapatkan informasi yang akurat tentang karakteristik dan objek yang diteliti, maka perlu alat ukur yang baik (sahih) yakni alat ukur yang mempersyaratkan beberapa hal, sehingga alat ukur tersebut menghasilkan informasi yang mengandung ketetapan yang tinggi, dan kesalahan kecil, sehingga hasilnya dapat diandalkan (Asmin, 2004). Persyaratan alat ukur pendidikan, menurut Cronbach (1990) meliputi kesahihan (*validitas*) yang diperoleh melalui korelasi sebuah tes dengan suatu kriteria tes yang ditentukan, dan keterandalan (*reliabilitas*) alat ukur yakni suatu proses yang dilakukan oleh pengguna tes dalam mengumpulkan bukti untuk mendukung inferensi yang dibuat berdasarkan skor tes.

Menurut teori tes klasik kesahihan meliputi kesahihan isi, konstrak, dan kriteria (Crokcer & Algina, 1986). Validitas dapat berarti sejauh mana ketepatan dan kecermatan suatu alat ukur dalam melakukan fungsi ukurnya. Menurut Djaali (2000) bahwa validitas tes tinggi apabila tes tersebut menjalankan fungsi ukur secara tepat, atau memberikan hasil ukur yang sesuai dengan maksud dilakukannya pengukuran tersebut. Selanjutnya, reliabilitas artinya sejauh mana hasil pengukuran dapat dipercaya. Suatu hasil pengukuran hanya dapat dipercaya apabila dalam beberapa kali pelaksanaan pengetesan terhadap kelompok subyek yang sama diperoleh hasil yang relatif sama.

Pada pengukuran klasik ciri yang unik diperlihatkan dari kenyataan bahwa kelompok butir tes atau kelompok angket (kuesioner) tidak dapat dipisahkan dari kelompok peserta tes atau kelompok yang mengisi angket. Artinya, kelompok butir tes/ angket (kuesioner) yang sama harus dijawab oleh kelompok peserta tes yang sama. Jika kelompok tes yang sama dijawab kelompok peserta uji tes yang berbeda maka ciri karakteristik kelompok butir itu akan berubah, sehingga taraf kesukaran dan daya pembeda kelompok butir tes itu akan berubah semata-mata karena kelompok butir tes tersebut ditanggapi oleh kelompok peserta yang berbeda. Menurut Setiadi (1998) bahwa dalam teori klasik, statistik soal, misalnya indeks kesukaran soal tergantung pada sampel pengikut ujian. Kalau tes tersebut dikerjakan oleh siswa yang pandai maka soal-soal itu sepertinya mudah atau tingkat kesukaran soalnya menjadi besar, dan sebaliknya kalau dikerjakan oleh siswa yang kurang pandai maka soal itu sepertinya sukar atau tingkat kesukaran soal menjadi kecil. Jadi, soal-soal itu tidak konsisten atau berubah-ubah tergantung pada kemampuan kelompok sampel siswa yang menempuh ujian.

Sejalan dengan itu, jika kelompok peserta tes yang sama menjawab kelompok butir tes yang berbeda maka ciri kelompok peserta akan berubah. Dalam hal ini kemampuan atau sikap para peserta berubah semata-mata karena peserta tes yang menjawab butir tes yang berbeda, sehingga kelompok peserta yang sama dan kelompok butir tes yang berbeda akan menunjukkan ciri peserta yang berbeda.

Pada penskoran klasik ada keterkaitan antara kedua kelompok butir tes dan kelompok peserta tes, yang memungkinkan munculnya beberapa hal: 1) kelompok peserta uji tes yang cirinya diskor perlu mengikuti tes yang sama pada saat yang bersamaan, sehingga perlu dihindari kebocoran butir tes sebelum tes dilaksanakan; 2) keterkaitan antara kelompok butir dan kelompok peserta tes mengakibatkan tafsiran skor diarahkan pada kelompok peserta tes yang menjawab tes tersebut. Biasanya tafsiran tersebut mengacu ke acuan norma; dan 3) tes yang terlalu mudah atau terlalu sukar tidak akan mencerminkan kemampuan peserta tersebut dengan akurat, sehingga kedua bentuk tes tersebut dipertimbangkan untuk diganti.

Responden memiliki kemampuan θ yang biasanya berbeda di antara responden. Butir memiliki taraf sukar butir b yang biasanya berbeda di antara butir. Pada pengukuran terjadi pertemuan di antara kemampuan responden dengan tara sukar butir. Jawaban atau tanggapan responden terhadap butir membuahkan hasil ukur. Dalam hal tertentu, hasil ukur menunjukkan salah atau betul. Pada skala dikotomi, jawaban salah sering diberi skor 0 dan jawaban betul diberi skor 1. Hasil ukur dapat juga dinyatakan dalam bentuk probabilitas jawaban betul (nilai dari 0 sampai 1). Probabilitas jawaban betul ditentukan oleh padanan di antara kemampuan responden dengan taraf sukar butir.

Probabilitas jawaban betul $Pgi(\theta)$ adalah probabilitas jawaban betul responden ke-g pada butir ke-i. Tidak selalu taraf sukar butir sepadan dengan kemampuan responden. Butir terlalu mudah atau

terlalu sukar tidak dapat menunjukkan kemampuan responden, sehingga akurasi pengukuran menjadi rendah. Kecocokan di antara kemampuan responden dengan taraf sukar butir menghasilkan akurasi pengukuran yang tinggi. Kecocokan di antara kemampuan responden dengan taraf sukar butir menghasilkan akurasi pengukuran tertinggi melalui ketentuan:

$$P(\theta) = P_{min} + 0{,}5\ (P_{maks} - P_{min})$$

Karena peluang menjawab benar atau Pmaks = 1 maka ketentuan ini menjadi:

$$P(\theta) = P_{min} + 0{,}5\ (1 - P_{min})$$

Pencocokan di antara kemampuan responden dengan taraf sukar butir dapat dilakukan jika mereka independen. Jika taraf sukar butir (b) independen dari kemampuan (θ) maka dapat dicari nilai taraf sukar butir yang cocok dengan kemampuan (θ).

Pada teori klasik, taraf sukar butir bergantung (*dependent)* kepada kemampuan responden. Bagi responden berkemampuan tinggi, butir menjadi tidak sukar (mudah). Bagi responden berkemampuan rendah, butir menjadi sukar. Pada butir tidak sukar (mudah), tampak kemampuan responden menjadi tinggi. Pada butir sukar, tampak kemampuan responden menjadi rendah. Taraf sukar butir bergantung kepada kemampuan responden. Butir yang sama akan terasa berat bagi mereka yang berkemampuan rendah dan terasa ringan bagi mereka yang berkemampuan tinggi.

Kemampuan responden bergantung kepada taraf sukar butir. Mereka yang mengerjakan butir sukar akan tampak berkemampuan rendah sedangkan mereka yang mengerjakan butir mudah akan tampak berkemampuan tinggi. Teori pengukuran klasik (teori ujian klasik) tidak dapat digunakan untuk pencocokan kemampuan responden dengan taraf sukar butir (karena mereka dependen). Pada teori klasik, terdapat interdependensi di antara kemampuan responden dan taraf sukar butir. Sebaiknya cara penyebutan hasil pengukuran disandingi dengan nama alat ukur. Misalnya, 450 TOEFL, 630 SPMB.

Untuk mengatasi kelemahan pada pengukuran klasik, penggunaan pengukuran modern ditampilkan yakni untuk menganulir ketidakterpisahan antara kelompok peserta tes dengan kelompok butir tes. Artinya, prinsip pengukuran modern adalah penetapan ciri butir, walaupun ciri peserta tes berbeda. Dengan kata lain, ciri dari kelompok butir adalah tetap walaupun dijawab peserta tes yang berbeda. Dengan demikian, berlaku pula bahwa ciri peserta akan tetap sama, walaupun mereka menjawab butir tes yang berbeda. Secara luas pembahasan tentang pengukuran modern dikaji secara mendalam dalam teori responsi butir.

Teori Responsi Butir (*Item Response Theory*)

Teori Responsi Butir (*Item Response Theory disingkat IRT*) dinamai juga sebagai Teori Ciri Laten (*Latent Trait Theory disingkat LTT*) atau Lengkungan Karakteristik Butir (*Item Characteristic Curve disingkat ICC*). Untuk memudahkan pengertian, di sini hanya digunakan istilah IRT. Seperti disebutkan di atas, pada hakekatnya IRT bertujuan untuk mengatasi kelemahan yang terdapat pada pengukuran klasik. Pada IRT, peluang jawaban benar yang diberikan siswa, ciri atau parameter butir, dan ciri atau parameter peserta tes dihubungkan melalui suatu model formula yang harus ditaati baik oleh kelompok butir tes maupun kelompok peserta tes (Hambleton & Rogers, 1991). Artinya, butir yang sama terhadap peserta tes yang berbeda harus tunduk pada aturan rumus itu, atau peserta tes yang sama terhadap butir tes yang berbeda juga harus patuh terhadap rumus tersebut. Dalam proses semacam ini terjadilah apa yang disebut *invariansi* di antara butir tes dan peserta tes. Pada pengukuran modern, taraf sukar butir tidak dikaitkan langsung dengan kemampuan responden.

Perbedaan mendasar antara pengukuran klasik dengan pengukuran modern terletak pada *invariansi* penskoran, di mana penskoran modern adalah invarians (tidak berubah atau tetap) terhadap butir tes serta terhadap peserta tes. Menurut Lord (1990) bahwa invariansi parameter-parameter butir tes melalui kelompok peserta tes merupakan karakteristik yang paling penting dari IRT. Kita biasanya memikirkan bahwa indeks kesukaran butir tes sebagai proporsi jawaban yang benar sehingga sukar untuk membayangkan bagaimana indeks kesukaran tes dapat menjadi invariant terhadap kelompok peserta tes dari tingkat kemampuan yang berbeda.

Pada pengukuran modern, taraf sukar butir dikaitkan langsung dengan karakteristik butir. Taraf sukar butir pada pengukuran modern terletak pada

: $P(\theta) = Pmin + 0,5 (P_{maks} - P_{min}) = P_{min} + 0,5 (1 - P_{min})$. Pada pengukuran modern, taraf sukar butir langsung dikaitkan dengan karakteristik butir. Kemampuan tinggi dan rendah memiliki taraf sukar butir yang sama. Kemampuan responden dan taraf sukar butir menjadi independen. Pengukuran modern dapat digunakan untuk pencocokan kemampuan responden dengan taraf sukar butir.

Teori responsi butir perlu menentukan model karakteristik butir yang digunakan. Model karakteristik butir dapat berbentuk satu parameter (1P), dua parameter (2P), tiga parameter (3P), atau model lain. Di sini pembahasan dibatasi pada satu sampai tiga parameter serta pada sekor dikotomi, yaitu: *1P : $P(\theta) = f(b, \theta)$ 2P : $P(\theta) = f(a, b, \theta)$ dan 3P : $P(\theta)$* = (a, b, c, θ). Satu, dua, dan tiga adalah banyaknya parameter butir. Parameter θ adalah parameter kemampuan responden. Parameter b adalah parameter taraf sukar butir. Pada 1P dan 2P, $b = \theta$ ketika $P(\theta) = 0,5$. Pada *3P,* $b = \theta$ ketika $P(\theta) = 0,5 (1 + c)$. Parameter a adalah parameter daya beda butir. Parameter c adalah parameter terkaan betul jawaban butir.

Tujuan Responsi Butir

Teori responsi butir membebaskan responden dan butir dari interdependensi, sehingga. taraf sukar butir tidak lagi bergantung kepada kemampuan responden. Kemampuan responden tidak lagi bergantung kepada taraf sukar butir. Melalui independensi di antara taraf sukar butir dan kemampuan responden, dapat dipilih butir yang cocok dengan responden. Dalam hal terjadi kecocokan di antara taraf sukar butir dan kemampuan responden, maka: kalau taraf sukar butir diketahui, kemampuan responden dapat ditentukan. Kalau kemampuan responden diketahui, taraf sukar butir dapat ditentukan.

Proporsi jawaban benar di dalam sebuah kelompok peserta tes tidak secara nyata mengukur kesulitan tes tersebut. Proporsi tersebut tidak hanya menjelaskan butir tes tetapi juga kelompok peserta yang dites. Ini merupakan suatu tujuan dasar untuk kesepakatan analisis statistik butir tes, yang dikenal dengan istilah invariansi. Yang menjadi dasar invariansi adalah taraf sukar butir tidak langsung dikaitkan dengan kemampuan responden melainkan dikaitkan dengan lengkungan karakteristik butir pada persamaan : $P(\theta) = P_{min} + (1 - P_{min})$

Misalkan suatu butir memiliki parameter butir a_1 = 1,27 dan b_1 = – 0,39. Butir ini diberikan kepada responden dengan kemampuan agak rendah dan dari mereka diperoleh lengkungan dengan a_1 = 1,27 dan b = – 0,39. Butir yang sama diberikan kepada responden dengan kemampuan agak tinggi dan dari mereka diperoleh lengkungan dengan a_1 = 1,27 dan b_1 = – 0,39. Pada responden dengan kemampuan agak rendah. Melalui perhitungan pada data diperoleh lengkungan dengan b_1 = - 0,39. Terlihat bahwa dua hasil ini adalah sama.

Asumsi Teori Reponsi Butir

Dalam teori responsi butir taraf sukar butir dan daya beda butir tes tetap sama, walaupun butir tes tersebut diselesaikan oleh kelompok peserta tes yang berbeda. Untuk itu, teori responsi butir mengembangkan model yang menghubungkan parameter butir dengan kemampuan peserta tes. Menurut Hambleton (1991) asumsi untuk model teori responsi butir secara mendalam digunakan, sehingga hanya satu kemampuan yang diukur dengan butir-butir tes tersebut. Hal ini dinamakan *unidimensi.* Suatu konsep yang menghubungkan keunidimensian adalah apa yang disebut dengan independensi lokal (*local independence*) yang akan didiskusikan berikutnya.

Asumsi lain dalam model teori responsi butir adalah fungsi karakteristik yang secara khusus melukiskan hubungan antara variabel kemampuan yang tidak teramati dengan variabel kemampuan yang teramati. Asumsi-asumsi tersebut juga menyangkut karakteristik butir tes yang relevan terhadap kinerja peserta tes pada suatu butir tes tersebut. Perbedaan besar antara model-model *Item Response Theory* dalam pemakaian bersama adalah dalam jumlah dan tipe serta karakteristik-karakteristik yang diasumsikan untuk kinerja peserta tes. Jadi dalam teori responsi butir dengan asumsi-asumsi tersebut, maka dalam setiap soal harus diwakili oleh satu *Item Characteristic Curve* (ICC). *Item Characteristic Curve* adalah pernyataan Matematika yang berhubungan dengan probabilitas keberhasilan peserta tes sesuai dengan kemampuannya.

Unidimensi

Asumsi unidimensi terpenuhi apabila butir-butir di dalam perangkat tes hanya mengukur satu kemam-

puan peserta tes. Misalnya butir-butir yang termuat di dalam perangkat tes bertujuan untuk mengukur kemampuan peserta tes dalam mata pelajaran Matematika. Butir-butir yang dikonstruksi berupa soal cerita dan berbentuk dikotomi. Apabila peserta tes memberi respon yang salah maka tidak dapat diketahui apakah kesalahan itu disebabkan oleh ketimpangan peserta tes pada mata pelajaran Matematika atau bahasa. Dalam kenyataannya sulit mendapatkan suatu butir yang mengukur hanya satu kemampuan peserta tes.

Menurut Dali S Naga (1992) bahwa persyaratan unidimensi ditujukan untuk mempertahankan invariansi pada teori responsi butir. Kalau butir tes sampai mengukur lebih dari satu dimensi, maka jawaban terhadap butir itu merupakan kombinasi dari berbagai kemampuan peserta tes. Akibatnya, tidak lagi diketahui kontribusi dari setiap kemampuan terhadap jawaban peserta tes tersebut. Dengan mengganti butir tes atau kelompok peserta tes, tidak dapat lagi dipertahankan invariansi pada ukuran ciri butir tes dan pada ukuran ciri peserta tes, sehingga ketidakmampuan mempertahankan syarat invariansi ini akan bertentangan dengan tujuan teori responsi butir tersebut.

Dengan terpenuhinya persyaratan unidimensi tersebut maka diperlukan cara untuk menentukan apakah suatu butir tes merupakan unidimensi atau tidak. Untuk hal ini, maka digunakan metode analisis faktor. Dalam hal ini penggunaan analisis faktor bertujuan untuk memperlihatkan pada kelompok faktor mana butir itu berada. Setiap faktor hanya menunjukkan suatu dimensi indikator tes. Dengan demikian, setiap dimensi indikator tes terhimpun dalam satu faktor yang melibatkan beberapa butir tes yang diperlukan, Faktor-faktor tersebut mungkin meliputi motivasi, kecemasan, kemampuan bekerja cepat, kecenderungan menebak bila dalam keadaan ragu-ragu menjawab, dan keterampilan kognitif di dalam menjumlahkan, serta faktor dominan lain yang diukur dengan sehimpunan butir tes (Asmin, 2004).

Independensi Lokal

Asumsi independensi lokal dibagi menjadi dua yaitu independensi lokal terhadap respons peserta tes dan independensi lokal terhadap butir tes (James J. Allen & Yen, 1989). Independensi lokal terhadap respons peserta tes, memiliki arti bahwa betul salahnya peserta tes menjawab sebuah butir tidak terpengaruh oleh betul salahnya peserta tes yang lain dalam menjawab butir tersebut. Sedangkan indepensi lokal terhadap butir, memiliki arti bahwa betul salahnya seorang peserta tes menjawab sebuah butir tidak terpengaruh oleh betul salahnya peserta tes yang lain dalam menjawab butir yang lain.

Ada independensi lokal responden terhadap butir dan ada independensi lokal butir terhadap responden. Pada peserta tes di lokasi yang sama, probabilitas menjawab betul P(θ) untuk butir berbeda adalah independen satu terhadap lainnya. Misalkan responden yang memiliki kemampuan yang sama mengerjakan butir X1, X2, X3, …, XN, maka sesuai dengan rumus independensi pada probabilitas, berlaku

$$P(X_1 I X_2 I X_3 \ldots\ldots I X_N) =$$

$$P(X_1)P(X_2)P(X_3)\ldots\ldots P(X_N) \text{ atau}$$

$$P(X_1 I X_2 I X_3 \ldots\ldots I X_N) = \prod_{i=1}^{i=N} P(X_i)$$

$$QP(X_i) = 1 - P(X_i)$$

Independensi lokal butir terhadap responden. Pada butir di lokal yang sama, probabilitas men-jawab betul P(θ) untuk responden berbeda adalah independen satu terhadap lainnya. Independensi lokal dapat diuji dengan dua cara, yaitu: secara eksak melalui rumus probabilitas, dan secara statistika melalui uji ketergantungan khi-kuadrat.

Pengujian Melalui Rumus Probabilitas

Independensi lokal tercapai apabila data memenuhi rumus independensi pada probabilitas. Berikut contoh pengujian melalui rumus probabilitas: Responden mengerjakan butir ke-1 dan ke-2 dengan probabilitas jawaban

		Butir ke-2		
		1	0	
Butir	1	0,086	0,420	0,506
ke-1	0	0,083	0,411	0,494
		0,169	0,831	1

Apakah terdapat independensi lokal? Berdasarkan data di atas maka perhitungan probabilitasnya adalah sebagai berikut:

P(11)=0,086 P1(1)P2(1) = (0,506)(0,169) = 0,086
P(10)=0,420 P1(1)P2(0) = (0,506)(0,831) = 0,420

P(01)=0,083 P1(0)P2(1) = (0,494)(0,169) = 0,083
P(00)=0,411 P1(0)P2(0) = (0,494)(0,831) = 0,411

Jadi, terdapat kecocokan sehingga mereka adalah independen secara lokal.

Pengujian secara Statistika

Pengujian dilakukan pada taraf signifikansi tertentu melalui hipotesis: H_0: ada independensi lokal. H_1: tidak ada independensi lokal. Distribusi probabilias pensampelan adalah distribusi probabilias khi-kuadrat dan statistik uji c^2 adalah:

		Butir ke-2		
		1	0	
Butir	1	A	B	A+B
ke-1	0	C	D	C+D
		A+C	B+D	N

Statistik uji adalah menggunakan persamaan berikut:
$X^2=N(AD-BC)^2/(A+B)(C+D)(A+C)(B+D)$
dengan $\nu = I\,N =$ banyaknya responden, dan A, B, C, D = frekuensi. Dengan kriteria pengujian adalah: Tolak H0 jika $\chi 2 > \chi 2(\alpha)(\nu)$. Terima H0 jika $\chi 2 \leq \chi 2(\alpha)(\nu)$.

Prinsip independensi lokal dinyatakan oleh asumsi bahwa secara formal, probabilitas (sukses pada butir *i* yang diberikan θ) sama dengan probabilitas (sukses pada butir *i* yang diberikan q dan juga diberikan kinerjanya pada butir *j, k, …*). Jika $u_i = 0$ atau 1 menyatakan sekor butir ke-i, maka dapat ditulis dengan:

$$P(u_i = 1/\theta) = P(u_i = 1/\theta, u_f, u_k \ldots)$$

Menurut Lord (1990) secara matematika pernyataan indepensi lokal berarti bahwa probabilitas sukses seluruh butir tes sama dengan perkalian dari bagian-bagian probabilitas sukses tersebut. Sebagai contoh, ada tiga butir tes *i, j, dan k,* maka :

$$P(u_i = 1, u_j = 1, u_k = 1/\theta) =$$
$$P(u_i = 1/\theta)P(u_j = 1/\theta)P(u_k = 1/\theta)$$

Independensi lokal menginginkan setiap dua butir tidak berkorelasi apabila θ adalah tetap. Secara definisi tidak diinginkan butir-butir tidak berkorelasi dalam kelompok, dimana θ bervariasi. Dalam hal tertentu, independensi lokal secara otomatis mengikuti keunidimensian.

Menurut Crocker dan Algina (1986), dalam teori responsi butir secara bersama-sama digunakan konsep-konsep yang lebih umum terhadap keterikatan dan kebebasan statistik untuk menyatakan tentang hubungan antara varaiabel-variabel. Untuk dua sekor butir dikotomi konsep-konsep tersebut dapat diilustrasikan secara numerik sebagai berikut. Bila diketahui responsi dari 40 responden pada suatu butir soal hasil akhirnya adalah seperti Tabel 1.

Atau peluang jawaban tersebut dibentuk seperti Tabel 2.

Tabel 2. Peluang Jawaban Butir 1 dan Butir 2

	1	0	
1	0,100	0,200	0,300
0	0,500	0,200	0,700
	0,600	0,400	

Dari tabel 2 tersebut dapat dihitung besar perkalian setiap peluang sebagai berikut:

P(11) = 0,10 $P_1(1)P_2(1) = (0,30)(0,60) = 0,18$
P(10) = 0,20 $P_1(1)P_2(1) = (0,30)(0,40) = 0,12$
P(01) = 0,50 $P_1(0)P_2(1) = (0,40)(0,60) = 0,24$
P(00) = 0,20 $P_1(0)P_2(1) = (0,70)(0,04) = 0,28$

Dari hasil perkalian peluang-peluang tersebut dapat disimpulkan bahwa tidak terdapat independensi lokal, karena tidak memenuhi syarat independensi lokal (Nitko, 1992).

Keempat kondisi persamaan tersebut mengatakan bahwa skor-skor butir adalah bebas jika masing-masing peluang susunan jawaban untuk kedua butir sedemikian rupa sehingga peluang pada ruas kiri dari persamaan dapat dihitung dengan mengetahui hanya peluang jawaban benar dan salah untuk masing-masing butir tersebut. Dengan demikian, dapat disimpulkan bahwa sebuah tes adalah unidimensional jika butir-butir tes tersebut secara statistik adalah tidak bebas di dalam populasi yang dilibatkan.

Tabel 1. Responsi jawaban siswa sejumlah 40 responden

Butir	Responsi responden							
1	00000	11000	00011	00010	00100	00000	11001	10101
2	01100	00011	10000	11111	11111	11100	00110	01111

Invarian

Seperti disebutkan di atas, pada hakikatnya *Item Response Theory* (IRT) bertujuan untuk mengatasi kelemahan yang terdapat pada pengukuran klasik. Perbedaan mendasar antara pengukuran klasik dengan pengukuran modern terletak pada *invariansi* pensekoran, di mana pensekoran modern adalah invarians (tidak berubah) terhadap butir tes serta terhadap peserta tes. Menurut Lord (1990: 126) bahwa invariansi parameter-parameter butir tes melalui kelompok peserta tes merupakan karakteristik yang paling penting dari IRT. Dapat dikatakan bahwa indeks kesukaran butir tes sebagai proporsi jawaban yang benar sehingga sukar untuk membayangkan bagaimana indeks kesukaran tes dapat menjadi invarian terhadap kelompok peserta tes dari tingkat kemampuan yang berbeda.

Dalam IRT, proporsi jawaban benar, ciri (parameter) butir, dan ciri peserta dihubungkan melalui rumus, di mana muncul masalah dalam menentukan rumus responsi butir atau rumus karakteristik butir yang dikenal sebagai penentuan model responsi butir atau model karakteristik butir. Masalah lainnya adalah bagaimana menentukan nilai parameter butir dan nilai parameter peserta yang diistilahkan sebagai pengestimasian parameter, baik parameter butir maupun parameter peserta, yang disebut sebagai pengkalibrasian butir. Untuk pemeriksaan hasilnya dilakukan estimasi parameter, yang bertujuan sebagai pencocokan model.

Karakteristik Butir

Karakteristik butir dalam teori responsi butir terdiri dari daya beda butir, taraf sukar butir dan faktor kebetulan menjawab betul pada butir dinyatakan berturut-turut dengan huruf a, b, dan c. Parameter peserta tes adalah kemampuan peserta tes yang dinyatakan dengan θ. Kemampuan peserta tes terhadap butir ke-j dinyatakan dalam bentuk probabilitas jawaban betul $P_j(\theta)$. Skor respon-den mencerminkan kemampuan res-ponden sehingga skor responden dan kemampuan responden merupakan parameter responden. Kemampuan responden merupakan suatu kontinum dari rendah ke tinggi. Skor responden tinggi menunjukkan kemampuan tinggi dan skor responden rendah menunjukkan kemampuan responden rendah.

Taraf Sukar Butir

Pada umumnya makin mudah butir atau makin kecil b, maka makin besar probabilitas responden untuk menjawab butir itu dengan benar sehingga nilai $P(\theta)$ menjadi besar. Sebaliknya makin sukar butir atau makin besar b, maka makin kecil probabilitas responden menjawab butir itu dengan benar sehingga nilai $P(\theta)$ menjadi kecil. Dengan demikian mudah sukarnya suatu butir menurut Dali S. Naga (1998: 34) sering dikaitkan dengan kemampuan responden dengan taraf sukar butir yakni dengan $(\theta - b)$.

Ada butir yang sukar, ada butir yang sedang, dan ada butir yang mudah. Taraf sukar butir merupakan suatu kontinum dari mudah ke sukar. Taraf sukar butir ke-i dinyatakan dengan bi. Makin tinggi taraf sukar butir bi, diperlukan kemampuan responden θ yang makin tinggi untuk dapat menjawabnya dengan betul, jika θ>bi maka Pi (θ) tinggi, sedangkan jika θ<bi maka Pi (θ) rendah. Untuk mendapatkan hasil analisis yang baik, seharusnya jumlah soal paling tidak 40 sampai dengan 50 dan jumlah peserta tes paling tidak 400 orang.

Kontinum taraf sukar berimpit dengan kontinum kemampuan responden. Taraf sukar butir adalah peluang untuk menjawab benar suatu soal pada tingkat kemampuan tertentu yang umumnya dinyatakan dalam bentuk indeks. Indeks tingkat kesukaran ini pada umumnya dinyatakan dalam bentuk proporsi yang besarnya berkisar 0,00 – 1,00. Soal yang memiliki indeks 0,00 artinya tidak ada siswa yang menjawab benar, indeks 1,00 artinya siswa menjawab benar butir tes.

Perhitungan indeks tingkat kesukaran ini dilakukan untuk setiap nomor soal. Pada prinsipnya sekor rata-rata yang diperoleh peserta didik pada butir soal yang bersangkutan dinamakan tingkat kesukaran butir soal itu. *Tingkat Kesukaran adalah jumlah siswa yang menjawab benar butir soal dibagi dengan jumlah siswa yang mengikuti tes.* Fungsi tingkat kesukaran butir soal pada umumnya dihubungkan dengan tujuan tes (Aiken, 1994). Misalnya untuk ujian semester digunakan butir soal yang memiliki tingkat kesukaran sedang, untuk keperluan seleksi digunakan butir soal yang memiliki tingkat kesukaran tinggi atau sukar, dan untuk keperluan diagnostik maka digunakan butir soal yang memiliki tingkat kesukaran rendah atau mudah.

Semakin besar indeks tingkat kesukaran yang diperoleh dari hasil perhitungan, berarti semakin mudah soal itu. Probabilitas jawaban betul pada butir ke-i berhubungan dengan letak θ terhadap bi atau terhadap $(\theta - bi)$ atau $Pi(\theta) = f(\theta - b)$. Ini dikenal sebagai kararteristik butir satu parameter $Pi(\theta) = f(\theta, bi)$. Nilai taraf sukar butir ke-i ditentukan oleh $\theta - bi = 0$ atau $bi = \theta$ pada saat $Pi(\theta) = 0,5$.

Suatu butir dikatakan mudah atau sukar bergantung dari kemampuan peserta tes. Apabila kemampuan peserta tes lebih dari taraf sukar butir maka dapat dikatakan butir itu mudah dan sebaliknya apabila kemampuan peserta tes kurang dari taraf sukar butir maka dapat dikatakan bahwa butir itu sukar. Tingkat kesukaran butir soal dapat mempengaruhi bentuk distribusi total sekor tes. Untuk tes yang sangat sukar (TK<0,25) distribusinya berbentuk positif skewed, sedangkan tes yang mudah (TK>0,8) distribusinya berbentuk negatif *skewed.*

Taraf sukar butir mempunyai dua kegunaan, yaitu kegunaan bagi guru dan kegunaan bagi pengujian dan pengajaran (Nitko, 1996). Kegunaan bagi guru adalah: 1) sebagai pengenalan konsep terhadap pembelajaran ulang dan memberi masukan kepada siswa tentang hasil belajar mereka; dan 2) memperoleh informasi tentang penekanan kurikulum atau mencurigai terhadap butir soal yang bias. Adapun kegunaannya bagi pengujian dan pengajaran adalah: 1) pengenalan konsep yang diperlukan untuk diajarkan ulang; 2) tanda-tanda terhadap kelebihan dan kelemahan pada kurikulum sekolah; 3) memberi masukan kepada siswa; 4) tanda-tanda kemungkinan adanya butir soal yang bias; dan 5) merakit tes yang memiliki ketepatan data soal.

Tingkat kesukaran butir soal juga dapat digunakan untuk memprediksi kemampuan peserta didik oleh pendidik. Misalnya satu butir soal termasuk kategori mudah, maka prediksi terhadap informasi ini adalah: 1) pengecoh butir soal itu tidak berfungsi; dan 2) sebagian besar peserta didik menjawab benar butir soal itu; artinya bahwa sebagian besar peserta didik telah memahami materi yang ditanyakan. Analisis secara klasik ini memiliki keterbatasan, yaitu tingkat kesukaran sangat sulit untuk mengestimasi secara tepat karena estimasi tingkat kesukaran dibiaskan oleh sampel. Di samping kedua kegunaan tersebut, dalam konstruksi tes, taraf sukar butir sangat penting karena taraf sukar butir dapat: 1) mempengaruhi karakteristik distribusi sekor (mempengaruhi bentuk dan penyebaran sekor tes atau jumlah soal dan korelasi antar soal); dan 2) berhubungan dengan reliabilitas, semakin tinggi korelasi antar soal semakin tinggi reliabilitas (Dali S. Naga, 1998). Demikian pula semakin tinggi nilai reliabilitas butir tes, semakin tinggi pula validitas butir soal tersebut.

Daya Beda Butir

Ada butir yang memiliki ciri: dapat dijawab dengan betul oleh kebanyakan responden yang berkemampuan tinggi, tidak dapat dijawab dengan betul oleh kebanyakan responden yang berkemampuan rendah. Butir demikian memiliki daya untuk membedakan responden berdasarkan kemampuan mereka. Butir memiliki parameter berupa daya beda butir. Daya beda butir adalah kemampuan suatu butir soal dapat membedakan antara peserta didik atau warga belajar yang telah menguasai materi yang ditanyakan dan warga belajar atau peserta didik yang belum menguasai materi yang ditanyakan.

Dengan kata lain daya beda butir adalah kemampuan suatu butir soal yang dapat membedakan antara siswa yang telah menguasai materi yang ditanyakan dan siswa yang belum menguasai materi yang ditanyakan. Jika tes atau soal mengukur hal yang sama, dapat diharapkan bahwa setiap peserta tes mampu menjawab soal dengan benar dan yang tidak mampu akan menjawab salah. Tingkat kesukaran berpengaruh langsung pada daya pembeda soal. Jika setiap orang menjawab benar (p=1), atau jika setiap orang menjawab salah (p=0), maka soal tidak dapat digunakan untuk membedakan kemampuan peserta tes (Surapranata, 2004). Manfaat daya beda butir adalah: 1) untuk meningkatkan mutu setiap soal melalui data empiriknya. Berdasarkan indeks daya beda butir, setiap butir soal dapat diketahui apakah butir soal itu baik, direvisi, atau tidak; dan 2) untuk mengetahui seberapa jauh setiap butir soal dapat mendeteksi atau membedakan kemampuan siswa, yaitu siswa yang telah memahami atau belum memahami materi yang diajarkan guru.

Apabila suatu butir soal tidak dapat membedakan kedua kemampuan siswa itu, maka butir soal itu dapat dicurigai kemungkinannya seperti berikut: 1) kunci jawaban butir soal itu tidak tepat; 2) butir soal itu memiliki dua atau lebih kunci jawaban yang benar; 3) kompetensi yang diukur tidak jelas; 4) pengecoh

tidak berfungsi; 5) materi yang ditanyakan terlalu sulit, sehingga banyak siswa yang menebak; dan 5) sebagian besar siswa yang memahami materi yang ditanyakan berpikir ada yang salah informasi dalam butir soalnya.

Indeks daya beda butir juga dinyatakan dalam bentuk proporsi. Semakin tinggi indeks daya beda butir berarti semakin mampu butir yang bersangkutan membedakan siswa yang telah memahami materi dengan siswa yang belum memahami materi. Indeks daya beda berkisar antara -1,00 sampai dengan +1,00. Semakin tinggi daya beda butir tes, maka semakin baik butir tes tersebut. Jika daya beda butir negatif berati lebih banyak kelompok bawah (peserta didik yang tidak memahami materi) menjawab benar butir tes dibanding dengan kelompok atas (peserta didik yang memahami materi yang diajarkan guru di kelas).

Untuk menggambarkan tentang daya beda butir maka dibuat grafik yang menunjukkan kemiringan kurva. Kecuraman pada lengkungan merupakan koefisien arah a pada fungsi a(θ-b). Makin curam makin besar koefisien arah a. Pada butir ke-i, daya beda butir dinyatakan sebagai koefisien arah yang menunjukkan kecuraman pada lengkungan yakni ai sehingga Pi (θ) = f (ai (θ-bi)). Selain itu indeks daya beda juga bisa dihitung dengan korelasi point biserial maupun korelasi biserial. Kelebihan korelasi point biserial: 1) memberikan refleksi kontribusi soal secara sesungguhnya terhadap fungsi tes. Maksudnya adalah mengukur bagaimana baiknya butir berkorelasi dengan kriterion; 2) sederhana dan langsung berhubungan dengan statistik tes; dan 3) tidak pernah mempunyai value 1,00 karena hanya variabel-variabel dengan distribusi bentuk yang sama yang dapat berkorelasi secara tepat, variabel kriterion dan skor dikotomi tidak mempunyai bentuk yang sama. Indeks daya pembeda dihitung atas dasar pembagian kelompok menjadi dua bagian, yaitu kelompok atas yang merupakan kelompok peserta tes yang berkemampuan tinggi dengan kelompok bawah yaitu kelompok peserta tes yang berkemampuan rendah. Kemampuan tinggi ditunjukkan dengan perolehan skor yang tinggi dan kemampuan rendah ditunjukkan dengan perolehan skor yang rendah (Messick, 1989).

Indeks daya pembeda didefinisikan sebagai selisih antara proporsi jawaban benar pada kelompok atas dengan proporsi jawaban benar pada kelompok bawah (Surapranata 2004). Adapun kelebihan korelasi biserial (Millman & Greene, 1993) adalah: 1) cenderung lebih stabil dari sampel ke sampel; 2) penilaian lebih akurat tentang bagaimana butir tes dapat diharapkan untuk membedakan pada beberapa perbedaan point di skala abilitas; dan 3) value koefisien korelasi biserial yang sederhana lebih langsung berhubungan dengan indikator diskriminasi *Item Characteristic Curve (ICC).*

Tingkat Kebetulan Betul pada Butir

Ada kalanya butir itu berbentuk pilihan ganda sehingga responden yang tidak memiliki kemampuan pun masih mungkin menjawab benar melalui terkaan. Dalam bentuk probabilitas, katakan saja bahwa tingkat kebetulan pada jawaban benar adalah c, maka untuk butir ini, probabilitas jawaban benar karena kebetulan adalah $P(\theta) = c$. Kalau jumlah pilihan ganda itu adalah empat (misalkan A, B, C, D), maka melalui terkaan saja terdapat 1 di antara 4 kemungkinan bahwa jawaban itu benar.

Dalam hal ini probabilitas jawaban benar karena kebetulan adalah ¼ atau 0,25 sehingga c = 0,25 (Dali S. Naga, 1998). Pada butir pilihan ganda dapat saja terjadi bahwa jawaban betul dicapai melalui terkaan. Jawaban betul ini adalah kebetulan betul. Tingkat kebetulan menjawab betul pada butir ke-i dinyatakan dengan parameter butir ci dan merupakan probabilitas jawaban betul minimum. Secara keseluruhan kita mengenal tiga karakteristik butir, yaitu a, b, dan c. Di samping itu, responden memiliki satu karakteristik yakni kemampuan responden. Karakteristik ini juga dikenal sebagai satu parameter pada karakteristik responden.

Pi(θ) min = ci. Di sini, taraf sukar butir bi tidak diperoleh melalui probabilitas jawaban betul *Pi(θ) = 0,5* melainkan pada : *Pi(θ) = ci + 0,5 (1- ci) = 0,5 (1 + ci)*. Bentangan *Pi (θ)* tidak lagi dari 0 sampai 1,0 melainkan dari ci sampai 1,0 yakni selebar (1-ci) sehingga: *f (ai (-θ bi))* menjadi *(1 ci) f (ai(-θ bi))* dan probabilitas jawaban betul menjadi: *Pi (θ) = ci + (1 - ci) f (ai (θ - bi))*. Di sini terdapat tiga parameter butir ai, bi, dan ci sehingga dikenal sebagai karakteristik butir tiga parameter dengan persamaan: *Pi (θ) = f (θ, ai, bi, ci).*

Penyusunan Tes Hasil Belajar Akhir Ujian Nasional

Penyusunan Tes Hasil Belajar Akhir baik secara

lokal maupun Nasional perlu dilakukan secara terencana dan teratur. Ujian Akhir Nasional dilakukan dengan skala yang lebih besar yang dilaksanakan setiap tahun, di mana soal-soal yang diberikan telah tersimpan dalam Bank Soal sehingga memudahkan untuk diakses dalam memenuhi kebutuhan tes Ujian Akhir Nasional yang setiap saat dapat diambil bila diperlukan.

Menurut Kumaidi (2000) untuk mengembangkan suatu tes dan sejumlah butir soal yang *defensible* maka prosedur pengembangan perlu ditradisikan, dalam arti proses pengembangan tes (dan penulisan butir soal) dimulai dengan pengembangan rancangan atau kisi-kisi tes, yang didahului oleh pembedahan kurikulum yang memuat segala informasi tentang tes tersebut. Rancangan tes ini memuat tujuan penilaian yang akan dilakukan, tempo (waktu yang ditempuh) untuk pelaksanaan pengujian, pesan utama kurikulum (sasaran pembelajaran dan garis besar topik materi uji), indikator butir soal (ciri-ciri penguasaan materi uji dan pencapaian sasaran pembelajaran), serta jumlah dan bentuk butir soal (per-indikator, per topik, dan keseluruhan tes). Sebaran butir soal dalam tes seharusnya memperhatikan keseimbangan tuntutan penguasaan sesuai dengan pesan kurikulum, sehingga memberi nuansa keterwakilan topik bahasan.

Menurut Jihad (2010), ada sembilan langkah yang harus ditempuh dalam mengembangkan tes hasil atau prestasi belajar, yaitu: 1) menyusun spesifikasi tes; 2) menulis soal tes; 3) menelaah soal tes; 4) melakukan uji coba tes; 5) menganalisis butir soal; 6) memperbaiki tes; 7) merakit tes; 8) melaksanakan tes; dan 9) menafsirkan hasil tes. Khusus mengenai uji coba tes, dalam penyusunan tes untuk mengukur prestasi hasil pembelajaran yang diselenggarakan oleh guru di kelas seperti ulangan harian, ulangan umum, dan ulangan kenaikan kelas, tidak harus dilakukan secara tersendiri. Pembakuan tes dilakukan melalui beberapa kali ujicoba. Sedangkan Djaali (2004) menjelaskan bahwa, penyusunan dan pengembangan tes dimak-sudkan untuk memperoleh tes yang valid, sehingga hasil ukurnya dapat mencerminkan secara tepat hasil belajar yang dicapai oleh masing-masing individu peserta tes setelah selesai mengikuti pembelajaran. Adapun langkah-langlah kontruksi tes yang ditempuh adalah sebagai berikut: 1) menetapkan tujuan tes; 2) analisis kurikulum; 3) analisis buku pelajaran dan sumber materi belajar lainnya; 4) membuat kisi-kisi; 5) penulisan tujuan instruksional khusus; 6) penulisan soal; 7) telaah soal (*face validity*); 8) reproduksi tes terbatas; 9) uji coba tes; 10) analisis hasil uji coba; 11) revisi soal, dan l2) merakit soal menjadi tes.

Langkah awal dalam mengembangkan tes adalah menetapkan spesifikasi tes, yaitu berisi uraian yang menunjukkan keseluruhan karakteristik yang harus dimiliki suatu tes. Spesifikasi yang jelas akan mempermudah dalam menulis soal, dan siapa saja yang menulis soal akan menghasilkan tingkat kesulitan yang relatif sama. Penyusunan spesifikasi tes mencakup kegiatan berikut ini: 1) menentukan tujuan tes; 2) menyusun kisi-kisi tes; 3) memilih bentuk tes; dan 4) menentukan panjang tes (Setiadi, 2009).

Selanjutnya, menurut Setiadi (1998) menyatakan bahwa setiap tahun soal-soal yang digunakan harus dibuat oleh suatu panitia khusus yang dibentuk untuk keperluan ujian nasional, sehingga setiap tahun harus dikeluarkan dana yang besar untuk keperluan revisi soal-soal tersebut. Untuk keperluan keamanan juga diperlukan beberapa alternatif paket tes *(paralel form)*, di mana soal-soal pada suatu paket dengan paket yang lain dianggap sama tingkat kesukaran soalnya hanya karena dianggap dibuat berdasarkan pada kisi-kisi yang sama tanpa didasarkan pada data empirik hasil uji coba soal di lapangan.

Pengembangan rancangan tes ini melibatkan spesialis (termasuk guru) bidang studi, sehingga bila rancangan tes telah selesai disusun maka rancangan tes tersebut harus divalidasi, melalui penelaahan pakar dan teman sejawat, sehingga benar-benar sesuai dengan pesan kurikulum. Untuk mengatasi variasi butir soal yang berlebihan, dengan pemahaman indikator butir soal, ada baiknya dikembangkan apa yang disebut oleh Nitko (1992) sebagai spesifikasi butir soal (*item specification*). Spesifikasi ini menyangkut uraian tentang batasan dan rambu-rambu yang harus dipatuhi oleh penulis butir soal.

Gronlund (1985) menyarankan beberapa hal dalam pengkonstruksian tes, diantaranya: 1) stem item tersebut sebaiknya memaknai butir itu sendiri dan menampilkan masalah tertentu; 2) stem butir tes melibatkan banyak kemungkinan jawaban dan bebas dari materi yang tidak relevan; 3) gunakan pernyataan stem butir yang bersifat negatif *hanya* ketika hasil belajar yang dikehendaki cukup berarti

(signifikan); 4) Semua alternatif jawaban secara gramatikal konsisten dengan stem butir tersebut; 5) sebuah butir secara jelas hanya mengandung satu jawaban benar terbaik; 6) butir-butir tes digunakan untuk mengukur pemahaman yang mengandung beberapa hal baru, tetapi harus berhati-hati; 7) semua pengecohan harus masuk akal; 8) assosiasi verbal antara stem dan jawaban yang benar harus dihindarkan; 9) secara relatif, panjang pilihan jawaban tidak menunjukkan suatu petunjuk untuk jawaban tersebut; 10) jawaban benar sebaiknya muncul pada masing-masing posisi pilihan atas beberapa kesamaan pendekatan, tetapi dalam urutan random; 11) gunakan dengan hemat pilihan-pilihan khusus seperti tidak satu pun jawaban di atas benar atau semua jawaban di atas benar; dan 12) jangan gunakan butir-butir pilihan berganda ketika butir yang lainnya lebih tepat.

Suatu tes harus mengukur hasil belajar dalam skala yang sama dan pendekatan yang mungkin dilakukan antara lain: 1) pemakaian butir soal penjangkar (*common items*) untuk beberapa set tes; 2) pemakaian butir soal yang telah terkalibrasi (butir soal yang diketahui karakteristiknya pada satu skala umum); dan 3) kombinasi kedua pendekatan itu yakni soal penjangkar dipilih dari butir yang terkalibrasi (Kumaidi, 2000). Dalam hal ini peranan IRT cukup berguna untuk menyamakan skala tersebut. Setelah soal-soal berkualitas terpilih berdasarkan *professional adjustment* dari para ahli bidang studi dan ahli pengukuran (*measurement specialist*) dan juga didukung data empirik hasil uji coba soal, maka kegiatan berikutnya adalah membuat skala dan menentukan di mana setiap soal terletak dalam skala tersebut (Setiadi, 1998).

Menurut Naga (1992) dari waktu ke waktu bank butir terus mengalami pengembangan dengan pemasukan butir-butir baru serta peniadaan butir-butir usang. Dalam penelitian digunakan teori skor modern. Untuk membentuk perangkat soal yang baik dibutuhkan banyak hal, terutama dari aspek esensial yang membutuhkan pengkajian lebih mendasar dan mendetail baik ditinjau dari kacamata pengukuran klasik maupun pengukuran modern sehingga pemanfaatan tes dapat menghasilkan fungsi informasi butir tes maupun fungsi informasi ujian yang cukup tinggi. Karenanya tidak ada satu tes yang sempurna, selama berbagai persyaratan yang telah diuraikan di atas belum seluruhnya dipenuhi.

Simpulan dan Saran

Simpulan

Pada hakikatnya proses pengukuran semuanya baik. Hanya kekonsistenan pelaksana dan penilai hasil pengukuran di samping kejujuran memberi penilaian adalah yang utama. Kecanggihan alat ukur modern belum tentu bermanfaat bagi peserta didik, selama hal itu dilakukan setengah hati. Teori responsi butir atau *item response theory* merupakan alternatif pilihan yang bertujuan melepaskan diri dari ketergantungan tes yang diberikan dengan sampel peserta tes. Dalam hal ini walaupun soal-soal tersebut dikerjakan oleh siswa yang pandai atau siswa yang kurang pandai, indikasi tingkat kesukaran suatu soal tetap tidak berubah.

Untuk mengukur kemampuan peserta tes yang sangat beragam di Indoensia, seperti Ujian Nasional, seharusnya digunakan juga ujian atau tes yang berbeda tingkat kesukaran soalnya, supaya adil dan juga akurat hasilnya. Peserta tes atau ujian (seperti Ujian Nasional) yang mengerjakan tes atau ujian yang berbeda tingkat kesukaran soalnya, tetap bisa dibandingkan kemampuannya, asalkan soal-soal dalam ujian tersebut berasal atau diambil dari bank soal yang sudah dikalibrasi dengan konsep *item response theory*.

Kekhawatiran dengan ketidaklulusan perlu disikapi secara wajar oleh semua pihak, khususnya sekolah dengan memperbaiki proses pembelajaran. Apabila upaya perbaikan proses pembelajaran telah dilakukan, sesungguhnya tidak ada sesuatu yang perlu dikhawatirkan, karena seluruh bahan ujian sudah mengacu pada kurikulum yang berlaku. Kelemahan-kelemahan yang ada dalam pelaksanaan Ujian Nasional perlu diidentifikasi dan dijadikan sebagai masukan dalam perbaikan pelaksanaan Ujian Nasional ke depan, dalam rangka membangun suatu sistem ujian akhir yang handal, yang dapat memberikan informasi akurat bagi pembangunan pendidikan.

Saran

Keseragaman penerapan tes secara nasional perlu dipertimbangkan lebih arif, mengingat tingkat kemampuan yang beragam sesuai lingkungan tempat tinggal peserta tes. Walaupun penerapan kurikulum berlaku secara nasional, namun faktor lingkungan tempat sekolah juga perlu dipertimbangkan. Konsep utama teori responsi butir adalah adanya kesesuaian

tingkat kesukaran suatu tes dengan kemampuan siswa yang menjawab adalah sesuatu yang tidak dapat diabaikan. Harus diingat bahwa nilai a (daya pembeda soal) yang tinggi, dan nilai c (tebakan jawaban) yang rendah, tanpa dibarengi nilai b (tingkat kesukaran soal) yang mendekati kemampuan *(θ)* akan memberikan nilai fungsi informasi butir tes yang rendah.

Pemanfaatan program komputer dalam menganalisis hasil tes sudah saatnya digunakan terutama untuk mengatasi berbagai kesalahan yang mungkin dilakukan secara manual, sehingga akurasi hasil analisis dapat dipertanggung jawabkan. Untuk masa yang akan datang disarankan Ujian Nasional sudah dapat melaksanakan ujian dengan sistem individual tes dengan menggunakan *Computer Adaptive Test* (CAT). Dengan menggunakan CAT permasalahan-permasalahan yang dihadapi dalam pelaksanaan tes secara kelompok klasikal seperti yang dilaksanakan dalam Ujian Nasional sekarang ini dapat dihindari.

Soal-soal Ujian Nasional harus dikembangkan berdasarkan bank soal yang sudah dikalibrasi dengan konsep teori responsi butir. Pada akhirnya keberhasilan siswa tidak hanya ditentukan oleh faktor hasil ujian hasil belajar saja, akan tetapi faktor-faktor lain, seperti kerajinan, kehadiran, hasil ujian bulanan, pengerjaan pekerjaan rumah, dan faktor-faktor lain seharusnya menjadi pertimbangan lain dalam menentukan kelulusan peserta didik.

Pustaka Acuan

Asmin. 2004. Implementasi Teori Responsi Butir dan Fungsi Informasi Butir Tes dalam Pengujian Hasil Belajar Akhir di Sekolah. *Jurnal Pendidikan dan Kebudayaan,* X (48): 234-245.

Azwar, Saifuddin. 2001. *Tes Prestasi. Fungsi Pengembangan Pengukuran Prestasi Belajar.* Yogyakarta: Pustaka Pelajar Offset.

Aiken, Lewis R. 1988. *Psychological Testing and Assessment.* Boston: Allyn and Bacon, Inc.

Crocker, Linda, & Algina, James. 1986. *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston, Inc.

Cronbach, Lee J. 1990. *Essentials of Psychological Testing.* New York: Harper Collins Publishers.

Dali S. Naga. 1998. Karakteristik Butir pada Alat Ukur Model Dikotomi, *Arkhe: Jurnal Ilmiah Psikologi,* III (4): 34-42.

Dali, S. Naga. 1992. *Pengantar Teori Sekor Pada Pengukuran Pendidikan*. Jakarta: Besbats.

Djaali. 2004. *Pengukuran Dalam Bidang Pendidikan*. Jakarta: Program Pascasarjana Universitas Negeri Jakarta.

Gronlund, Norman. E. 1985. *Measurement and Evaluation in Teaching*. New York: Macmillan Publishing Company.

Hambleton, Ronald K; Swaminathan, H; dan Jane Rogers, H. 1991. *Fundamentals of Item Response Theory.* London: SagePublications.

Jihad, Asep, Abdul Haris. 2011. *Evaluasi Pembelajaran.* Multi Pressindo: Yogyakarta.

Kumaidi. 2000. Standardisasi Butir Soal. *Jurnal Pendidikan dan Kebudayaan.* V (5): 132-143.

Lord, Frederick, M.1990. *Aplications of Item Response Theory to Practical Testing Problems.* New Jersey: LawrenceErlbaum Associates, Publishers.

Mary J.Allen and Wendy M Yen, 1989, *Introduction to Measurement Theory,* California: Broke.

Nitko, Anthony. J. 1992. *Criterion Reference Testing Workshop*: *Handouts and Reading Material Tidak dipublikasikan).* Cipayung, Bogor: Examination Development Unit (Puslitbang Sisjian).

Nitko, Anthony J. 1996. *Educational Assessment of Student,* Second Edition. Ohio: Merrill an Imprint of Prentice Hall Englewood Cliff.

Messick, S. 1989. *Educational Measurement,* 3rd edition, New York: Macmillan.

Millman, Jason and Greene, Jennifer. 1993. The Spesification and Development of Tests of Achievement and Ability in Robert L. Lin (Editor), *Educational Measurement,* Third Edition. Phoenix: American Council on Education, series on Higher Education Oryx Press.

Peraturan Pemerintah Nomor 19 Tahun 2005 Tentang Standar Nasional Pendidikan

Setiadi, Hari. 1998. Bank Soal yang Dikalibrasi dengan Konsep IRT Memecahkan Permasalahan Ujian-ujian Sistematik yang Diadakan pada Periode-periode Tertentu, *Jurnal Kajian Dikbud* IV (13).

Setiadi, Hari. 2009. Permasalahan dan Solusinya dalam Pelaksanaan Ujian Nasional di Masa Mendatang, *Matahari: Jurnal Penelitian dan Pendidikan.*X (1): 66-74.

Surapranata, Sumarna. 2004. *Analisis, Validitas, Reliabilitas Dan Interpretasi Hasil Tes,* Rosdakarya: Bandung.

Wibowo, Mungin Eddy. 2011. Kondisi Psikologis Siswa dalam Menghadapi Ujian Nasional, *Buletin BNSP: Media Komunikasi dan Dialog Standar Pendidikan.* VI (1): 7-11.