

KARAKTERISTIK TES PROGRAM PENYETARAAN D2 PGSD UNIVERSITAS TERBUKA: Implementasi Model Rasch

*Oleh:
Rustam*

Abstrak

Penelitian ini bertujuan: (1) memperoleh informasi tentang karakteristik butir dan tes ujian akhir semester Program Penyetaraan D2 PGSD UT, berdasarkan analisis dengan pendekatan teori respons butir model Rasch; (2) mengetahui butir-butir soal yang cocok dengan model Rasch dan (3) mengetahui kualitas butir-butir soal hasil telaah para ahli bidang studi dengan menggunakan pedoman telaah yang standar.

Perangkat tes yang diteliti sebanyak empat matakuliah dengan rincian jumlah sampel: perangkat tes PPDG 2130 sebesar 3009 lembar jawab, PPDG 2332 sebesar 3897 lembar jawab, PPDG 2131 sebesar 3186 lembar jawab, dan PPDG 2531 sebesar 4959 lembar jawab; yang diambil dengan teknik *proportional cluster random sampling*. Data penelitian ini diperoleh melalui dokumentasi. Secara kuantitatif, data diolah dengan menggunakan program Rascal untuk kepentingan estimasi parameter butir, dan secara kualitatif menggunakan pedoman yang telah standar yang dilakukan oleh tiga orang penelaah.

Hasil analisis data secara kualitatif menunjukkan bahwa keempat perangkat tes tersebut memiliki kualitas cukup baik sebagai alat ukur, karena tidak satu butirpun yang ditolak, sedangkan butir yang baik masing-masing berjumlah: 30 butir (60%) untuk PPDG 2130, 34 butir (56,7%) untuk PPDG 2332, 46 butir (76,7%) untuk PPDG 2131, dan 48 butir (80%) untuk PPDG 2531; selebihnya butir yang perlu direvisi.

Hasil analisis data empiris, baik hasil estimasi parameter butir maupun kecocokan data terhadap model menunjukkan bahwa keempat perangkat tes memiliki kualitas cukup baik dengan tingkat kesukaran sedang dan rerata sebesar 0,000 skala *logits*. Butir-butir soal yang baik masing-masing berjumlah: 39 butir (78%) untuk PPDG 2130, 46 butir (77,7%) untuk PPDG 2332, 33 butir (55%) untuk PPDG 2131, dan 46 butir (77,7%) untuk PPDG 2531, selebihnya butir soal yang ditolak.

Berdasarkan temuan ini, Universitas Terbuka dan semua pihak yang terlibat dalam pengujian dan penilaian, dianjurkan untuk terlebih dahulu menganalisis baik secara kualitatif maupun secara kuantitatif sebelum butir atau tes tersebut digunakan, sehingga penilaian lebih objektif dan kesalahan pengukuran dapat dihindari.

Pendahuluan

Mahasiswa UT yang terdaftar sebagai mahasiswa aktif hingga April 1996 sebanyak 390.315 orang (Depdikbud, 1996). Mereka tersebar di 27 propinsi, dikelola dalam 32 Unit Program Belajar Jarak Jauh (UPBJJ). Oleh karena itu, pilihan yang diambil UT untuk mengevaluasi hasil belajar mahasiswa pada akhir semester, umumnya menggunakan tes bentuk objektif. Untuk mata kuliah tertentu, yang jumlah mahasiswanya kurang dari 100 orang menggunakan tes bentuk uraian.

Tes yang digunakan sebagai alat uji pada akhir semester di UT, pada umumnya dikembangkan oleh tenaga edukatif dari Universitas atau Institut yang menjadi mitra UT, seperti: Universitas Indonesia, Universitas Gadjah Mada, Universitas Airlangga, IKIP Jakarta, IKIP Bandung, IKIP Yogyakarta dan IKIP Malang. Dengan demikian, para pengembang butir soal tersebut latar belakangnya dapat dikatakan heterogen. Oleh karena itu, ada kemungkinan para pengembang butir soal tidak mengikuti kaidah-kaidah pengembangan butir soal yang baik, sehingga dapat dipertanyakan apakah butir-butir soal yang telah dikembangkan tersebut merupakan butir soal yang baik?

Rerata hasil perolehan mahasiswa secara sampel dari keseluruhan mata kuliah yang ada di UT, yaitu mata kuliah keahlian (MKK) pada program Penyetaraan D2 PGSD, sejak masa ujian tahun 1993 semester genap sampai masa ujian tahun 1995 semester gasal, berada dalam rentang 36, 57-71, 88 dalam skala 100 (Puskom UT, 1996). Khusus rerata hasil perolehan mahasiswa untuk mata kuliah Pendidikan Matematika dan Pendidikan Ilmu Pengetahuan Alam, berturut-turut sebesar 49,52 dan 53,31.

Kenyataan seperti dikemukakan di atas, memberikan gambaran bahwa perangkat tes atau butir-butir soal ujian akhir semester UT, termasuk di dalamnya perangkat tes program Penyetaraan D2 PGSD UT, sebagian besar sukar dikerjakan oleh mahasiswa. Apabila tes ujian akhir semester sukar dikerjakan oleh mahasiswa, maka tes tersebut cenderung tidak berfungsi sebagai alat ukur yang baik.

Ditinjau lebih jauh, perangkat tes yang digunakan pada ujian akhir semester di UT, tampaknya memang belum memenuhi persyaratan tes yang baik, karena pengembangannya baru melalui tiga tahapan, yaitu

penyusunan spesifikasi, penulisan dan penelaahan tes. Pengembangan tes yang hanya melalui tiga tahapan tersebut baru memungkinkan tes yang secara teori baik, tetapi belum dapat dikatakan apakah tes yang demikian itu pada praktiknya nanti juga baik (Suryabrata, 1987).

Dalam analisis tes menggunakan data empiris, secara garis besar terdapat dua jenis analisis, yaitu analisis butir soal berdasarkan model teori tes klasik dan analisis butir soal model teori tes modern atau teori respons butir. Perbedaan mendasar antara kedua teori itu terletak pada hasil analisis. Teori tes klasik bersifat *sample bound*, artinya hasil analisis hanya berlaku pada sampel yang digunakan, atau hasil analisis terikat pada karakteristik peserta tes. Sedangkan pada teori respons butir bersifat *sample free*, artinya hasil analisis tidak terikat sampel yang digunakan. Dengan kata lain, hasil analisis dapat berlaku umum (Hambleton, 1985), (Hambleton, Swaminathan, dan Roger; 1991),

Kondisi yang ada pada perangkat tes ujian akhir semester UT sebagaimana dijelaskan di atas, merupakan masalah yang perlu dikaji dan dicari penyelesaiannya. Oleh karena itu, peneliti memandang perlu untuk melakukan kajian terhadap perangkat tes ujian akhir semester UT, khususnya program Penyetaraan D2 PGSD. Kajian tersebut akan dilakukan dengan dua pendekatan, yaitu pendekatan kualitatif dan kuantitatif. Kajian ini menggunakan teori respons butir model Rasch, karena model ini memiliki parameter daya pembeda dan faktor tebakan menjawab benar diasumsikan homogen untuk setiap butir, formula matematika yang digunakan cukup sederhana, dan estimasi parameter butir dan parameter kemampuan peserta dilakukan secara terpisah (Hayat, 1994), sehingga dimungkinkan akan memperoleh sifat objektivitas dalam menentukan parameter butir dan peserta, serta mengurangi kesalahan dalam pengukuran.

Berdasarkan uraian di atas, masalah dalam penelitian ini dibatasi pada kualitas tes pada program Penyetaraan D2 PGSD UT masa ujian 1996 semester gasal, dengan rumusan masalah: Bagaimana kualitas tes ujian akhir semester program Penyetaraan D2 PGSD UT?

Manfaat Penelitian

Hasil penelitian ini diharapkan dapat memberikan masukan kepada:

Karakteristik Tes Program Penyetaraan D2 PGSD UT

1) UT mengenai kualitas perangkat tes hasil analisis kualitatif yang dilakukan oleh para ahli bidang studi dengan menggunakan pedoman standar; dan karakteristik butir soal berdasarkan teori respons butir model Rasch, khususnya butir-butir soal pada perangkat tes matakuliah yang digunakan pada program Penyetaraan D2 PGSD; selanjutnya, butir-butir soal yang telah diketahui karakteristiknya dan cocok dengan model Rasch dapat dimasukkan ke dalam Pusat Pengembangan Soal (PPS) FKIP UT; 2) semua pihak yang terlibat dalam pengukuran tentang implementasi teori respons butir, khususnya model Rasch.

Metode Penelitian

1. Populasi dan Sampel Penelitian

Populasi dalam penelitian ini adalah seluruh lembar jawab peserta ujian akhir semester program penyetaraan D2 PGSD UT. Lembar jawaban ujian tersebut merupakan hasil respons mahasiswa, yang tersebar di seluruh wilayah Indonesia terhadap beberapa mata kuliah yang diujikan pada masa ujian 1996 semester gasal.

Mengingat berbagai keterbatasan, perlu dilakukan penyampelan terhadap data yang digunakan. Sampel tersebut diperoleh dengan cara *proportional cluster random sampling*. Matakuliah yang dijadikan subjek penelitian sebanyak empat matakuliah, yaitu PPDG 2130, PPDG 2332, PPDG 2131 dan PPDG 2531. Ukuran sampelnya berturut-turut sebesar 3009 lembar jawab, 3897 lembar jawab, 3186 lembar jawab, dan 4959 lembar jawab.

2. Teknik Analisis Data

Analisis data dilakukan dengan dua pendekatan, yaitu pendekatan kualitatif dan pendekatan kuantitatif. Pendekatan pertama merupakan analisis data melalui penelaahan butir soal pada setiap perangkat tes. Penelaahan ini dilakukan oleh tiga orang penelaah secara terpisah. Cara menelaah ditentukan berdasarkan pedoman telaah butir soal yang dikeluarkan oleh Puslitbang Sisjian Depdikbud.

Analisis data berikutnya adalah analisis dengan pendekatan kuantitatif. Data yang dianalisis berupa respons peserta ujian akhir semester terhadap

perangkat tes yang diikutinya. Pada pendekatan ini, data dianalisis untuk menentukan karakteristik butir dan karakteristik setiap perangkat tes serta kecocokan data setiap butir terhadap model, dengan menggunakan program Rascal dari MicroCAT Versi 3.00 tahun 1988.

Hasil Penelitian

1. Hasil analisis kualitatif

Hasil analisis butir secara kualitatif terhadap empat perangkat tes yang dikaji, menunjukkan bahwa keempat perangkat tes tersebut cukup baik. Hal ini ditunjukkan oleh cukup banyak butir-butir soal yang baik tanpa revisi dan tidak ada butir soal yang tidak diterima pada setiap perangkat tes. Perangkat tes PPDG 2130 sebanyak 30 butir soal (60%), PPDG 2332 sebanyak 34 butir soal (56,7%), PPDG 2131 sebanyak 45 butir soal (73,3%), dan perangkat tes PPDG 2531 sebanyak 48 butir soal (80%). Butir-butir soal yang membutuhkan revisi, masing-masing sebanyak 20 butir (40%) untuk perangkat tes PPDG 2130, 26 butir soal (43,3%) untuk PPDG 2332, 15 butir soal (25%) untuk PPDG 2131, dan sebanyak 12 butir soal (20%) untuk perangkat tes PPDG 2531.

Butir-butir soal yang diterima tetapi harus direvisi pada setiap perangkat tes, disebabkan oleh adanya kelemahan yang merata pada bidang materi, konstruksi dan bahasa. Kelemahan dalam penulisan butir soal tersebut dapat saja terjadi, karena seorang penulis soal di samping harus menguasai teknik penulisan soal, juga harus menguasai beberapa kemampuan lain, yaitu keahlian dalam bidang studi yang diuji, keahlian dalam pengukuran, dan keahlian dalam membahasakan gagasan (Suryabrata, 1987). Keahlian tersebut tidak dikuasai semua oleh beberapa penulis soal program PGSD UT, sehingga ditemui perangkat tes yang butir-butir soalnya perlu direvisi pada beberapa bidang penulisan soal.

Bidang yang menyumbang kelemahan terbesar pada semua perangkat tes adalah bidang konstruksi soal, yaitu masing-masing 17 butir soal (34%) untuk PPDG 2130, 19 butir soal (31,7%) untuk PPDG 2332, 14 butir soal (23,3%) untuk PPDG 2131, dan delapan butir soal (13,3%) untuk PPDG 2531. Kemudian bidang materi, yaitu masing-masing dua butir soal (4%) untuk PPDG 2130, tujuh butir soal (11,7%) untuk PPDG 2332, Tujuh

Karakteristik Tes Program Penyetaraan D2 PGSD UT

butir soal (11,7%) untuk PPDG 2131, dan satu butir soal (1,7%) untuk PPDG 2531. Selanjutnya bidang bahasa, yaitu masing-masing satu butir soal (1,7%) untuk PPDG 2130 dan PPDG 2332, empat butir soal (6,7%) untuk PPDG 2131, dan delapan butir soal (13,3%) untuk PPDG 2531.

Jika diamati lebih jauh pada setiap perangkat tes, ditemukan butir-butir soal yang perlu direvisi yang bukan hanya direvisi dalam satu bidang, tetapi juga direvisi dalam bidang yang lain. Pada PPDG 2130 tidak ada butir soal yang direvisi yang lebih dari satu bidang. Pada PPDG 2332 sebanyak dua butir soal yaitu butir 23 dan butir 25. Butir 23 memerlukan revisi bidang materi dan konstruksi dan butir 50 memerlukan revisi bidang konstruksi dan bahasa. Pada PPDG 2131 sebanyak delapan butir soal, yaitu butir 4 dan 13 memerlukan revisi pada semua bidang; butir 1, 15, 16, dan butir 28 memerlukan direvisi bidang materi dan konstruksi; dan butir 5 dan butir 8 memerlukan revisi bidang konstruksi dan bahasa. Pada PPDG 2531 sebanyak lima butir soal, yaitu butir 13, 19, 22, 23, dan butir 39, kesemuanya memerlukan revisi bidang konstruksi dan bahasa.

Butir-butir soal yang direvisi bidang materi, umumnya belum berfungsi pengecohnya (kriteria 2), terutama pada PPDG 2332 dan PPDG 2131, masing-masing tujuh butir soal (11,7%), yaitu butir 2, 11, 16, 18, 23, 42, dan butir 47; dan enam butir soal (10%), yaitu butir 1, 4, 13, 15, 16, dan butir 28. Sedangkan pada PPDG 2130 dan PPDG 2531 masing-masing satu butir soal, yaitu butir 9 dan butir 24. Butir-butir soal yang belum berfungsi pengecohnya pada umumnya butir soal yang berbentuk angka. Hal ini terjadi karena penulis soal pada umumnya hanya memikirkan kunci jawaban butir soal dari pada pengecohnya (distraktor), tanpa memperhatikan dari mana asal angka tersebut.

Kemudian beberapa butir soal memiliki lebih dari satu kunci jawaban (kriteria 3). Pada masing-masing perangkat tes, masing-masing satu butir soal, kecuali PPDG 2332. Butir 21 untuk PPDG 2130, butir 2 untuk PPDG 2131, dan butir 24 untuk PPDG 2531. Hal ini dapat saja terjadi, karena penentuan alternatif soal tersebut disyaratkan jawaban yang paling tepat. Akibatnya, timbul kesulitan bagi penulis soal menentukan mana jawaban yang paling tepat di antara alternatif jawaban yang disajikan.

Ditinjau dari segi konstruksi, jumlah butir soal yang direvisi menduduki peringkat teratas dibandingkan dengan butir soal yang direvisi pada bidang materi dan bahasa. Bila dibandingkan banyak butir yang direvisi antar kriteria telaah dalam segi ini, jumlah butir yang direvisi hampir merata di semua kriteria dalam setiap perangkat tes, kecuali kriteria pada: (1) perangkat tes PPDG 2130, kriteria 11 (pilihan jawaban yang berbentuk angka tidak disusun berdasarkan urutan besar kecilnya angka) dan kriteria 7 (gambar/grafik/tabel/diagram dan sejenisnya pada soal tidak berfungsi dan jelas), yaitu masing-masing sembilan butir soal (18%) dan lima butir soal (10%). (2) perangkat tes PPDG 2332, kriteria 11 yakni sebanyak 13 butir soal (21,7%); (3) perangkat tes PPDG 2131, lebih banyak pada kriteria 8 (pilihan jawaban tidak homogen dan logis) dan kriteria 9 (panjang pilihan jawaban relatif tidak sama), yaitu masing-masing enam butir soal (10,0%) dan tiga butir soal (5,0%); (4) perangkat tes PPDG 2531, lebih banyak pada kriteria 4 (pokok soal tidak dirumuskan dengan singkat, jelas, dan tegas) dan kriteria 9 (panjang pilihan jawaban relatif tidak sama), masing-masing sebanyak enam butir soal (10%); dan tiga butir soal (5%).

Butir soal yang pilihan jawabannya berbentuk angka yang tidak disusun berdasarkan urutan besar kecilnya angka, pada umumnya soal tersebut dalam menyelesaikannya memerlukan perhitungan dan angka. Persoalan ini pada dasarnya hanyalah teknis penulisan soal saja. Karena bila dicermati lebih jauh, soal tersebut alternatif jawabannya homogen dan logis serta pengecohnya telah berfungsi dengan baik.

Butir-butir soal yang tidak memenuhi kriteria gambar/grafik/tabel/diagram dan sejenisnya pada soal jelas dan berfungsi, pada umumnya butir soal tersebut gambarnya tidak lengkap, misalnya ukuran tidak disertai dengan satuan. Selain itu, ukuran gambar tidak sesuai dengan pernyataan yang ada dalam pokok soal serta gambar terlalu kecil ukurannya. Persoalan ini dapat menimbulkan salah interpretasi bagi peserta tes yang merespons butir-butir soal tersebut.

Butir-butir soal yang tidak memenuhi kriteria pilihan jawaban homogen dan logis adalah butir-butir soal yang disebabkan oleh beberapa hal, antara lain pengecohnya tidak sesuai dengan susbtansi pokok soal dan ada beberapa pengecoh yang jelas salah. Hal ini terjadi, karena penulis soal kesulitan

an menemukan pengecoh yang lebih tepat, atau soal tersebut memang sulit untuk ditemukan alternatif jawaban yang tepat. Demikian juga butir-butir soal yang tidak memenuhi kriteria panjang pilihan jawaban relatif sama, memang sulit ditemukan alternatif jawabannya, sehingga antara alternatif jawaban yang satu dengan yang lain panjang kalimatnya tidak seimbang.

Ditinjau dari segi bahasa, secara keseluruhan sudah baik. Terutama perangkat tes PPDG 2130 dan PPDG 2332, masing-masing hanya satu butir soal yang perlu direvisi pada bidang ini. Butir 14 PPDG 2130 perlu direvisi pada kriteria 14 dan butir 50 PPDG 2531 perlu direvisi pada kriteria 13. Perangkat tes PPDG yang perlu direvisi, dua butir soal pada kriteria 13, dan masing-masing satu butir pada kriteria 14 dan kriteria 16, yaitu secara berurut butir 4 dan 13, 13, dan butir 14. Perangkat tes PPDG 2531 yang perlu direvisi, enam butir soal pada kriteria 13 dan satu butir soal pada kriteria 14, yaitu secara berurut butir 1, 9, 13, 19, 22 dan butir 23; dan 27.

Butir-butir soal yang tidak menggunakan bahasa yang sesuai dengan kaidah bahasa Indonesia disebabkan oleh beberapa hal, di antaranya penggunaan istilah yang tidak tepat; susunan kata dalam pokok soal kurang tepat; dan penggunaan kata sandang yang tidak tepat atau kalau dihilangkan tidak mengubah makna kalimat tersebut, sehingga terkesan pemborosan kata. Butir-butir soal yang tidak menggunakan bahasa yang komunikatif, penyebabnya adalah tidak menggunakan kata penghubung atau kalimatnya tidak lengkap. Kemudian kriteria yang lain yang belum terpenuhi terletak pada pilihan jawaban yang mengulang kata/kelompok kata yang sama, tetapi yang tidak memenuhi tersebut hanya satu butir soal. Pengulangan kata/kelompok kata yang sama pada pilihan jawaban sebenarnya dapat dihindari, dengan cara memindahkan kata/kelompok kata yang sama tersebut ke dalam pokok kalimat dengan tanpa menghilangkan makna kalimat tersebut.

2. Hasil analisis kuantitatif

Hasil analisis butir dari program Rascal yang menghasilkan parameter tingkat kesukaran butir menunjukkan bahwa, dari keempat perangkat tes yang masing-masing berjumlah 60 butir soal, kecuali perangkat tes PPDG

2130 berjumlah 50 butir soal, hampir seluruh butir soal pada keempat perangkat tes tersebut dapat dikategorikan sebagai alat ukur yang baik. Perangkat tes PPDG 2130 sebanyak 49 butir soal (98%), PPDG 2332 sebanyak 54 butir soal (90%), PPDG 2131 sebanyak 57 butir soal (95%), dan perangkat tes PPDG 2531 sebanyak 58 butir soal (96,7%).

Butir-butir soal yang ditolak/tidak baik adalah butir soal yang memiliki parameter tingkat kesukaran > 2 atau < -2 dalam skala *logits*. Butir-butir soal yang dapat diterima/baik adalah butir yang memiliki tingkat kesukaran antara ≥ -2 dan ≤ 2 dalam skala *logits* pada kurve normal (Hamblethton *at al.*, 1991; Hamblethton & swaminathan, 1985). Butir soal yang ditolak pada setiap perangkat tes jumlahnya bervariasi. Perangkat tes PPDG 2130 hanya satu butir soal, yakni butir 10. PPDG 2332 sebanyak enam butir soal, yaitu butir 2, 17, 20, 31, 35, dan butir 56. PPDG 2131 sebanyak tiga butir soal, yaitu butir 2, 30, dan butir 38. Perangkat tes PPDG 2531 sebanyak dua butir soal, yaitu butir 25 dan butir 50.

Menurut Hamblethton *at al.*, (1991), butir-butir soal di atas sebaiknya dibuang atau diganti, karena butir-butir soal tersebut sangat sukar atau sangat mudah. Butir-butir soal yang termasuk dalam kategori sangat sukar atau sangat mudah cenderung tidak dapat melakukan fungsi ukurnya, sehingga soal tersebut tidak dapat membedakan peserta tes yang berkemampuan tinggi dengan peserta tes yang berkemampuan rendah.

Rerata parameter tingkat kesukaran butir setiap perangkat tes sebesar 0,000 dalam skala *logits*, yang juga berarti bahwa parameter tingkat kesukaran tes tersebut sebesar 0,000 dalam skala *logits*. Tes yang memiliki tingkat kesukaran sebesar 0,000 dalam skala *logits*, dapat diartikan tes tersebut bertingkat kesukaran sedang. Bila dikaitkan dengan *Item Characteristic Curve (ICC)*, tes tersebut dapat dijawab oleh peserta yang memiliki kemampuan 0,000 dalam skala *logits* dengan peluang 50% benar.

Estimasi parameter butir melalui program Rascal, selain menghasilkan estimasi parameter tingkat kesukaran butir, juga menghasilkan harga *chi square* untuk setiap butir. Harga statistik ini digunakan untuk menguji kecocokan data dengan model yang dipakai. Ketidacocokan ini ditunjukkan oleh harga statistik *chi square* empiris butir yang jauh melebihi harga *chi square* teoritis.

Karakteristik Tes Program Penyetaraan D2 PGSD UT

Butir-butir soal yang tidak cocok terhadap model tersebut berurut sebagai berikut. Perangkat tes PPDG 2130 sebanyak 10 butir soal (20%), PPDG 2332 sebanyak 10 butir soal (16,7%), PPDG 2131 sebanyak 24 butir soal (40%), dan perangkat tes PPDG 2531 sebanyak 13 butir soal (21,7%).

Butir-butir soal yang tidak cocok terhadap model di atas, secara empiris merupakan butir-butir soal yang tidak berfungsi sebagaimana mestinya atau butir soal yang bias. Butir-butir yang tidak cocok terhadap model, dapat disebabkan oleh materi, konstruksi, atau bahasa butir soal yang tidak baik (Hayat, 1994). Sehingga butir soal tidak mengukur peserta tes atau hasil ukur tidak menggambarkan kemampuan peserta tes yang sebenarnya.

Di samping itu, yang menyebabkan soal menjadi tidak berfungsi atau bias adalah pelaksanaan ujian yang tidak/kurang dikontrol atau diawasi, sehingga memungkinkan peserta tes bekerja sama dalam mengerjakan tes. Kalau ini terjadi, dapat dikatakan butir atau bahkan tes tersebut tidak berfungsi sebagai alat ukur yang semestinya. Di UT, hal ini mungkin saja terjadi, karena sistem ujian di UT bersifat nasional yang membutuhkan tenaga pengawas ujian yang tidak sedikit, sehingga tidak menutup kemungkinan yang menjadi pengawas ujian tersebut kurang mempedulikan peserta yang diawasinya.

Dengan demikian, yang menyebabkan butir soal atau tes tidak berfungsi sebagaimana mestinya, bukan hanya semata-mata disebabkan oleh unsur internal seperti unsur materi, konstruksi, dan bahasa, tetapi dapat juga disebabkan oleh unsur eksternal dari butir soal atau tes seperti sistem pengawasan ujian.

Analisis kuantitatif menghasilkan dua hal yang berbeda, yaitu parameter tingkat kesukaran butir dan kecocokan data butir terhadap model. Kedua-duanya berfungsi untuk menentukan butir yang dapat diterima atau tidak sebagai alat ukur yang baik, sehingga di akhir bahasan ini dapat disimpulkan bahwa keempat perangkat tes yang dianalisis dapat dikatakan sebagai tes yang cukup baik, karena butir-butir yang ditolak relatif sedikit. Pada perangkat tes PPDG 2130 sebanyak 11 butir soal (22%), PPDG 2332 sebanyak 14 butir soal (23,3%); PPDG 2131 sebanyak 27 butir soal (45%), dan perangkat tes PPDG 2531 sebanyak 14 butir soal (23,3%).

Kesimpulan

Berdasarkan temuan hasil penelitian dan pembahasan dapat dikemukakan beberapa kesimpulan seperti berikut.

Analisis kualitatif yang dilakukan terhadap empat perangkat tes, yaitu perangkat tes PPDG 2130, PPDG 2332, PPDG 2131, dan perangkat tes PPDG 2531; masing-masing berjumlah 60 butir soal, kecuali perangkat tes PPDG 2130 berjumlah 50 butir soal; menghasilkan butir yang baik masing-masing berjumlah 30 butir soal (60%) untuk PPDG 2130, 34 butir soal (56,7%) untuk PPDG 2332, 46 butir (76,7%) untuk PPDG 2131, dan 48 butir soal (80%) untuk PPDG 2531. Sedangkan selebihnya, secara berturut-turut 40%, 43,3%; 23,3%; dan 20% merupakan butir-butir soal yang memerlukan revisi. Dengan demikian, keempat perangkat tes yang dianalisis tersebut secara umum memiliki kualitas yang cukup baik.

Analisis kuantitatif menghasilkan dua harga statistik berupa harga estimasi parameter tingkat kesukaran butir dan harga *chi square*. Keduanya secara terpisah digunakan untuk menyaring butir soal yang baik dan tidak baik secara empiris. Ditinjau dari harga estimasi parameter butir, yang memenuhi kriteria butir yang baik untuk masing-masing perangkat tes: PPDG 2130 berjumlah 49 butir soal (98%), PPDG 2332 berjumlah 54 butir soal (90%), PPDG 2131 berjumlah 57 butir soal (95%), dan PPDG 2531 berjumlah 58 butir soal (96,7%). Sedangkan rerata parameter butir/parameter tes untuk semua perangkat tes sebesar 0,000 dalam skala *logits*.

Ditinjau dari harga *chi square* kecocokan data butir terhadap model butir-butir yang cocok terhadap model pada setiap perangkat tes berjumlah: 40 butir soal (80%) untuk PPDG 2130, 50 butir soal (83,3%) untuk PPDG 2332, 36 butir soal (60%), dan 47 butir soal (78,3%). Secara keseluruhan hasil analisis empiris menunjukkan keempat perangkat tes memiliki kualitas yang cukup baik dengan tingkat kesukaran sedang. Butir-butir soal yang baik masing-masing perangkat tes berjumlah: 39 butir soal (78%) untuk PPDG 2130, 46 butir soal (77,7%) untuk PPDG 2332, 33 butir soal (55%) untuk PPDG 2131, dan 46 butir soal (77,7%) untuk PPDG 2531.

Daftar Pustaka

- Allen, M.J. and Yen, W.M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/CPC.
- Crocker, L. and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart and Winston, Inc.
- Dali S. Naga. (1992). *Pengantar Teori Sekor pada Pengukuran Pendidikan*. Jakarta: Gunadarma.
- Depdikbud. (1996). *Universitas Terbuka sebuah Inovasi dalam Pendidikan Tinggi*. Jakarta: Universitas Terbuka.
- Depdikbud. (1995). Bank soal. *Buletin Pengujian dan Penilaian*, halaman 3-7. Jakarta: Pusion, Balitbang, Depdikbud.
- Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991), *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publication.
- Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational Measurement*. New York: Macmillan. Publishing Company. hal. 147-200.
- Hambleton, R.K., Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer Academic Publisher.
- Hayat, B. (1995). Pengantar model Rasch. *Buletin Pengujian dan Penilaian*, halaman 11-16 . Jakarta: Pusion, Balitbang, Depdikbud.
- Hulin, C.L., Drasgow F. and Parson C.K. (1983). *Item Response Theory: Application to Psychological Measurement*. Homewood, IL: Dorsey Professional Series.
- Linn, R.L. (Ed.) (1989). *Educational measurement* (3rd ed.). New York: Macmillan Publishing Company.
- Lord, M.F. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: LEA, Publishers.
- Sukamto, (1992). *Teknik Penentuan Sampel (Seri Metodologi Penelitian Modul 05)*. Yogyakarta: FPTK IKIP Yogyakarta.
- Suryabrata, S. (1987). *Pengembangan Tes Hasil Belajar*. Jakarta: CV Raja Wali.
- _____. (1995). Penggunaan soal pilihan ganda dalam ujian. *Buletin Pengujian dan Penilaian*, halaman 12-15. Jakarta: Pusion, Balitbang, Depdikbud.
- Umar, J. (1991). *Pengantar Penilaian Pendidikan (makalah)*. Jakarta: Pusion, Balitbang, Depdikbud.