ANALISIS BUTIR DENGAN TEORI TES KLASIK DAN TEORI RESPONS BUTIR

Oleh:
Djemari Mardapi

Abstrak

Tujuan penelitian ini adalah untuk mengungkapkan tingkat konsistensi hasil analisis butir tes, estimasi kehandalan tes, dan estimasi kemampuan antara teori tes klasik dan teori respons butir.

Populasi penelitian ini adalah sekor hasil tes kuasa (*power test*), sedang cuplikannya adalah sekor hasil tes potensi belajar mahasiswa baru FPTK IKIP Yogyakarta angkatan 1992. Tes potensi belajar ini dirancang untuk seleksi masuk ke perguruan tinggi yang dikembangkan oleh Pusat Pengembangan Sistem Pengujian Balitbang Depdikbud. Teknik analisis data yang digunakan adalah paket program MicroCat.

Kesimpulan penelitian ini adalah: 1) jumlah butir yang ditolak menurut teori tes klasik lebih banyak dibanding menurut teori respons butir, 2) indeks konsistensi hasil analisis butir adalah rendah, 3) terdapat hubungan yang tinggi antara hasil estimasi kemampuan pada kedua teori tersebut yaitu, 0,99 dan 0,98 untuk perangkat 1 dan perangkat 2, 4) indeks kehandalan tes menurut teori tes klasik dan teori respons butir sama-sama tinggi, dilihat dari besarnya kesalahan baku pengukuran. Selanjutnya disarankan agar para pengembang tes untuk tingkat regional mencoba menggunakan teori trespons butir dalam me- ngembangkan tes, dan perlu penelitian yang sejenis dengan data yang berbeda dan dengan paket program yang baru.

Pendahuluan

Pada saat ini ada dua teori pengukuran yang berkembang dan banyak digunakan dalam merancang dan menganalisis alat ukur atau tes. Pertama adalah teori tes klasik dikembangkan sejak tahun 1940 dan telah digunakan secara luas, sedang teori yang ke dua adalah teori respons butir, berkembang setelah teknologi komputer berkembang. Teori yang ke dua ini menggunakan lebih banyak asumsi dibandingkan dengan teori yang pertama, namun dapat menyajikan informasi lebih banyak.

Alat ukur yang baik harus memiliki bukti kesahihan dan kehandalan. Untuk itu setiap alat ukur yang baku harus melalui proses validasi sebelum digunakan. Menurut Cronbach (1971) validasi alat ukur adalah suatu proses yang dilakukan para pengembang tes dalam mengumpulkan bukti untuk mendukung inferensi yang dibuat berdasarkan sekor tes. Kesahihan menurut teori tes klasik ada tiga yaitu, kesahihan isi, konstruk, dan kriteria (Crocker & Algina, 1986:217). Banyak formula yang dikembangkan untuk mencari besarnya indeks kehandalan suatu alat ukur (Feldt & Brennan, 1989). Tiap formula dikembangkan berdasarkan asumsi tertentu.

Suatu tes pada prinsipnya ingin menaksir besarnya kemampuan seseorang dengan tingkat kesalahan yang sekecil mungkin. Untuk itu dibutuhkan tes yang memiliki butir-butir yang baik, yaitu melalui analisis butir. Analisis butir pada prinsipnya menaksir besarnya parameter butir. Teori tes klasik dan teori respons butir keduanya dapat digunakan untuk menganalisis butir tes dan menaksir besarnya kemampuan seseorang. Kesalahan pengukuran pada kedua teori tersebut dinyatakan dengan kesalahan baku pengukuran. Besarnya kesalahan ini pada teori tes klasik diperoleh dari indeks kehandalan tes, sedang pada teori respons butir diperoleh dari fungsi informasi tes.

Besarnya estimasi kemampuan seseorang dan parameter butir ditentukan oleh teori pengukuran yang digunakan. Oleh karena itu perlu dilakukan penelitian tentang teknik analisis yang paling tepat dalam menguji mutu suatu alat ukur.

Rumusan masalah penelitian ini adalah: 1) seberapa jauh kesamaan dan perbedaan hasil analisis butir tes dengan teori tes klasik dan dengan teori respons butir?, 2) berapa besar hubungan hasil estimasi kemampuan antara teori tes klasik dan teori respons butir?, 3) berapa besar estimasi kehandalan sekor hasil tes?

Penelitian ini bertujuan untuk membandingkan hasil analisis butir antara teori tes klasik dan teori respons butir, serta konsistensi estimasi kemampuan dengan kedua teori tersebut.

Manfaat penelitian ini secara teoritis diharapkan bisa memberikan sumbangan pada pengembangan metodologi pengukuran, sedang secara praktis sangat bermanfaat dalam pemilihan teknik analisis yang paling tepat.

Menurut teori tes klasik, sekor tampak (X) terdiri dari sekor sebenarnya (T) dan sekor kesalahan (E). Ada dua asumsi dasar yang digunakan pada teori tes klasik yaitu tidak ada korelasi antara sekor yang sebenarnya dan sekor kesalahan, dan rerata kesalahan acak pengukuran sama dengan nol (Allen & Yen, 1979). Berdasarkan asumsi tersebut dikembangkan sejumlah formula untuk menghitung besanya indeks kehandalan tes.

Ada tiga parameter butir yang diestimasi, yaitu tingkat kesukaran, daya beda, dan dugaan. Tingkat kesukaran adalah proporsi peserta yang menjawab benar, sedang daya pembeda adalah hubungan antara sekor butir tes dengan sekor total dan dikenal sebagai korelasi point biserial.

Formula yang banyak digunakan untuk menghitung indeks kehandalan tes menurut teori tes klasik adalah Spearman-Brown, dan Cronbach-alpha (Crocker & Algina, 1986:137). Formula Spearman Brown menggunakan asumsi tes pararel, yaitu suatu tes dibelah menjadi dua yang memiliki rerata dan simpangan baku yang sama. Formula Cronbach-alpha berdasarkan pada bentuk Tau-ekivalen, yaitu suatu tes dapat dibelah dua atau lebih dan masing-masing tidak harus memiliki rerata dan simpangan baku yang sama.

Keunggulan teori tes klasik terletak pada kemudahan dalam pemahaman kosep dan penggunaannya, sehingga banyak digunakan. Kelemahan teori ini terletak pada hasil etimasi parameter butir yang tergantung pada karakteristik peserta tes dan hasil estimasi parameter kemampuan tergantung pada karakteristik butir. Tes yang mudah akan

menghasilkan estimasi kemampuan yang tinggi, demikian pula sebaliknya, tes yang sulit akan menghasilkan estimasi kemampuan yang rendah.

Teori respons butir memiliki tiga model, yaitu model satu parameter, dua parameter, dan tiga parameter (Hambleton & Swaminathan, 1986). Model satu parameter dikenal dengan model Rasch, yaitu yang berasumsi bahwa: 1) semua butir memiliki daya pembeda yang sama, dan 2) peluang menjawab butir benar bagi mereka yang memiliki kemampuan rendah sama dengan nol. Dengan kata lain semua kurve karakteristik butir-butir model ini adalah sejajar atau mendekati sejajar. Oleh karena itu parameter butir pada model Rasch adalah hanya tingkat kesulitan butir, sedang parameter daya pembeda dianggap sama, dan dugaan pseudo dianggap sama dengan nol.

Persamaan model satu-parameter yang dikenal dengan model Rasch dapat ditulis sebagai berikut:

$$P_i(0) = \frac{e^{Da(0-bi)}}{1 + e^{Da(0-bi)}} \qquad 1)$$

P_i (0) adalah peluang menjawab benar butir i, D = 1,7, 0 adalah kemampuan, dan b adalah tingkat kesukaran butir. Model dua parameter menggunakan asumsi bahwa peluang menjawab benar bagi mereka yang memiliki kemampuan rendah adalah nol, sehingga hanya ada dua parameter butir yang ditaksir, yaitu tingkat kesukaran dan daya pembeda. Pada tiga parameter tidak menggunakan asumsi tentang parameter butir, sehingga tiga parameter butir, yaitu tingkat kesukaran, daya pembeda, dan faktor dugaan, ketiganya ditaksir besarnya.

Dilihat dari kesederhanaannya, model satu parameter tampak paling sederhana, namun menggunakan asumsi yang lebih banyak. Sifat ini yang menjadi pertimbangan bagi Balitbang Depdikbud untuk menggunakan model satu parameter, dikenal dengan model Rasch, dalam mengembangkan jaringan pengujian di Indonesia.

Teori respons butir menggunakan istilah informasi untuk menyatakan kehandalan tes. Fungsi informasi sangat berguna untuk konstruksi tes, pemilihan butir, penilaian presisi pengukuran, komparasi sejumlah tes, dan penentuan bobot dalam penyekoran. (Hambleton & Swaminathan, 1985:101). Besarnya informasi butir tes tergantung pada daya pembeda, tingkat kesukaran, dan dugaan pseudo seperti pada persamaan 2. Besarnya informasi pada prinsipnya tergantung pada tingkat kemampuan peserta tes. Oleh karena itu untuk memperoleh informasi yang maksimum, tingkat kesulitan tes harus sesuai dengan tingkat kemampuan yang mengikuti tes.

Besarnya informasi butir tes dan informasi tes yang merupakan penjumlahan dari informasi semua butir dapat dilihat pada persamaan 2 dan persamaan 3. Besarnya kesalahan baku pengukuran atau estimasi dapat dihitung dengan formula 4.

$$I_i(0) = \frac{[P'_i(0)]^2}{P_i(0)\,[1 - P_i(0)]}, \qquad 2)$$

dan fungsi informasi tes adalah:

$$I(0) = I_i(0) \qquad 3)$$

sedang besarnya kesalahan baku estimasi kemampuan adalah:

$$SE(0) = \frac{1}{I(0)} \qquad 4)$$

P_i' adalah turunan pertama peluang menjawab benar, $I_i(0)$ adalah informasi butir, $I(0)$ adalah informasi tes, sedang $SE(0)$ adalah kesalahan baku estimasi.

Kelemahan utama dari teori respons butir terletak pada dua hal, yaitu penghitungan yang lebih komplek dan membutuhkan ukuran cuplik-

an yang besar. Namun karena penghitungan pada teori respons butir menggunakan paket program komputer, kelemahan diatas dapat diatasi. Hanya masalah pemahaman pada teori respons butir membutuhkan pengetahuan tentang matematika dan statistik.

Cara Penelitian

Wilayah generalisasi hasil penelitian ini adalah semua tes kuasa. Populasi penelitian ini adalah respons terhadap tes potensi belajar subtes kuantitatif yang dikembangkan oleh Balitbang Depdikbud Jakarta. Cuplikannya adalah data hasil tes potensi belajar perangkat 1 dan perangkat 2 subtes kuantitatif mahasiswa FPTK IKIP YOGYAKARTA angkatan tahun 1992 yang berupa respons terhadap 40 butir soal dalam waktu 40 menit. Tes ini sudah diuji coba tiga kali dan sudah digunakan untuk seleksi masuk ke perguruan tinggi swasta sebanyak dua kali.

Data respons mahasiswa terhadap tes dianalisis dengan menggunakan paket program komputer MicroCat yang mencakup analisis dengan teori tes klasik dan dengan teori respons butir model Rasch. Model teori respons yang dipilih adalah yang paling sederhana dan yang digunakan oleh Balitbang Depdikbud, yaitu model satu parameter atau dikenal dengan model Rasch.

Menurut teori tes klasik, besarnya tingkat kesukaran butir yang baik adalah 0,30 sampai 0,70 (Allen & Yen, 1979:121), sedang daya beda yang diterima minimum 0,20. Pengecoh dikatakan berfungsi apabila ada peserta ujian yang memilih tiap pengecoh.

Indeks kehandalan tes yang baik menurut Linn adalah minimum 0,70 (Linn, 1989). Formula Cronbach alpha (Mehrens & Lehmann:1984) digunakan untuk menghitung besarnya indeks kehandalan tes. Estimasi kemampuan seseorang dinyatakan dengan jumlah butir yang dijawab benar.

Analisis dengan teori respons butir meliputi tingkat kesukaran, estimasi kemampuan, dan fungsi informasi tes. Satuan untuk tingkat kesukaran, kemampuan, dan informasi tes adalah logit, yaitu log bilangan

alam dari suatu persamaan untuk masing-masing parameter (Wright & Stone, 1979).

Untuk mengetahui tingkat konsistensi estimasi jumlah butir yang baik atau yang tidak baik digunakan koefisien phi (Hinkle, Wiersma, & Jurs, 1979:101). Untuk mengetahui hubungan hasil estimasi kemampuan antara teori tes klasik dan teori respons butir digunakan korelasi produk momen.

Asumsi daya pembeda yang homogen dijadikan kriteria untuk menentukan suatu butir cocok atau tidak dengan model Rasch. Untuk ini digunakan statistik Khai kuadrad.

Hasil Penelitian dan Pembahasan

Berdasarkan kriteria tingkat kesukaran soal, daya pembeda, dan distribusi respons, jumlah butir soal yang dinyatakan tidak baik pada tes perangkat 1 ada 21 buah. Terbanyak soal ditolak oleh kriteria tingkat kesukaran, yaitu sebanyak 13 butir, kemudian karena daya pembeda sebanyak 10 butir, dan paling sedikit karena kriteria distribusi respons yaitu sebanyak 8 butir.

Jumlah butir yang tidak baik pada perangkat 2 ditinjau dari tingkat kesukaran, daya pembeda, dan distribusi respons adalah 26 buah. Terbanyak disebabkan oleh tingkat kesukaran sebanyak 20 butir, kemudian disusul karena distribusi respons sebanyak 19 butir, dan yang terkecil disebabkan daya pembeda sebanyak 4 butir.

Indek kehandalan tes perangkat 1 0,79, sedang untuk perangkat 2 adalah 0,85, dan kesalahan baku pengukuran adalah sebesar 2,57. Tes dengan indek kehandalan sebesar 0,79 dan 0,85 dapat dinyatakan baik.

Butir yang tidak cocok dengan model Rasch untuk perangkat 1 ada 8 butir soal, sedang untuk perangkat 2 hanya 4 butir. Besarnya informasi maksimum butir tes dalam satuan logit, yaitu sebesar $1/4D^2$.

Besarnya informasi tes maksimum adalah 28,90, dengan kesalahan baku pengukuran sebesar 0,186.

Tingkat konsistensi hasil analisis butir antara teori klasik dengan teori respons butir adalah -0,025 untuk tes perangkat 1, dan sebesar 0,07 untuk tes perangkat 2. Korelasi hasil estimasi kemampuan antara pendekatan teori klasik dan teori respons butir untuk perangkat 1 adalah sebesar 0,993, sedang untuk perangkat 2 adalah 0,977. Estimasi indek kehandalan menurut teori tes klasik adalah 0,79 untuk perangkat 1, dan 0,89 untuk perangkat 2.

Besarnya kesalahan baku pengukuran adalah 2,74 untuk perangkat 1, dan 2,62 untuk perangkat 2. Rasio kesalahan pengukuran terhadap rentang sekor tes adalah sebesar 0,069 dan 0,066 berturut-turut untuk tes perangkat 1 dan perangkat 2.

Untuk model Rasch, informasi tes maksimum untuk perangkat 1 dan perangkat 2, adalah 28,90, dan besarnya kesalahan baku pengukuran tes adalah 0,186. Rasio kesalahan terhadap rentang sekor dari -3.0 sampai + 3.0 adalah sebesar 0,031.

Pada tes perangkat 1 dan perangkat 2 lebih banyak butir yang ditolak menurut teori tes klasik dibanding menurut teori respons butir. Indeks konsistensi estimasi parameter butir juga rendah, -0,025 dan 0,07. Hal ini disebabkan oleh persyaratan yang dituntut oleh teori tes klasik lebih banyak dibanding dengan teori respons butir. Persyaratan butir yang baik menurut teori tes klasik meliputi tingkat kesukaran, daya pembeda, dan distribusi respons, sedang persyaratan yang dituntut oleh Rasch model adalah homogenitas besarnya daya pembeda.

Besarnya korelasi hasil estimasi kemampuan menurut teori tes klasik dan menurut teori respons butir sangat tinggi, yaitu 0,99 untuk perangkat 1 dan 0,98 untuk perangkat 2. Temuan ini menunjukkan bahwa estimasi kedua pendekatan tersebut adalah konsisten. Perbedaan besarnya estimasi kemampuan yang dinyatakan dengan sekor menurut teori tes klasik dan teori respons butir disebabkan perbedaan skala yang

digunakan. Untuk itu dapat digunakan sekor baku dengan rerata 100 dan simpangan baku 15.

Estimasi kesalahan pengukuran yang dinyatakan dengan rasio kesalahan baku pengukuran dan rentang sekor, untuk kedua teori ini adalah sama-sama rendah, yaitu 0,069, 0,066 untuk perangkat 1 dan perangkat 2, sedang pada teori respons butir adalah 0,031. Dengan demikian dapat dikatakan bahwa estimasi indeks kehandalan tes dapat dikatakan sama-sama tinggi pada kedua teori tersebut.

Kesimpulan dan Saran

Hasil analasis data dapat disimpulkan sebagai berikut: 1) jumlah butir yang ditolak menurut teori tes klasik lebih banyak dibanding menurut teori respons butir, 2) konsistensi hasil analisis butir antara teori tes klasik dan teori respons butir adalah sangat rendah, 3) Hubungan hasil estimasi kemampuan menurut teori tes klasik dan menurut teori respons butir cukup tinggi yaitu 0,99 dan 0,98, untuk perangkat 1 dan perangkat 2, 4) indeks kehandalan tes menurut teori tes klasik dan menurut teori respons butir sama-sama dinyatakan tinggi, demikian juga kesalahan baku pengukuran sama-sama dinyatakan rendah.

Keterbatasan penelitian ini terletak pada paket program versi lama yang di dalamnya tidak ada estimasi besarnya informasi tes, serta ukuran cuplikan yang digunakan relatif kecil, sehingga hasil estimasi bisa tidak stabil. Besarnya informasi tes dihitung secara manual dengan bantuan kalkulator.

Saran

1. Para pengembang tes untuk tingkat nasional atau regional sebaiknya mencoba menggunakan teori respons butir dalam mengembangkan suatu tes, khususnya model Rasch,
2. Perlu penelitian yang sejenis dengan menggunakan paket program yang lebih baru dan dengan data yang lain.

Daftar Pustaka

Birnbaum, (1968) A. Some latent trait models and their use in inferring an examinee's ability. Dalam F. M . Lord & M. R. Novick. *Statistical thoeries of mental test scores.* Reading, Mass: Addison Wesley.

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

Cronbach, L. J. (1971). Test validation. Dalam R.L. Thorndike (Ed.) *Educational measurement.(*2nd ed.) Washington DC: American Council on Education.

Hambleton, R. K ., & Cook, L. L. (1977) Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement,* 14, 2, h.75-96.

Hambleton, R. K., & Swaminathan, H. (1985). *Item rsponse theory: Principles and applications.* Boston: Kluwer-NIjhoff Publ.

Hambleton, R. K. & Van der Linden, W. J. (1982). Advances in item response theory. *Applied Psychological Measurement,* 1982,6, 4, h.373-378.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1979). *Applied statistics for the behavioral sciences.* New Jersey: Houghton Mifflin Company.

Linn, R. L. (1989). *Educational measurement.* New York:Mac Millan Publishing.

Lord, F. M. A. (1952). Theory of test scoress. *Psychometric Monograph,* No. 7.

Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, New Jersey:Lawrence Erlbaum Associates, Publishers.

Mehrens, W. A., & Lehmann, I. J. (1984). *Measurement and evaluation in Educational dan Psychology.* New York: Holt, Rinehart, Winston.

Wright, B, D. & Stone, M. H. (1979). *Best test design.* Chicago: Mesa Press.