PENSKALAAN BUTIR FORMAT RESPONS PILIHAN DAN RESPONS BEBAS BERDASARKAN MODEL *RASCH* DAN *PARTIAL CREDIT*

Oleh:
Ekohariadi

Abstrak

Penelitian melihat pengaruh jumlah parameter butir, kategori respons bebas (RB), pengaruh sampel terhadap akurasi estimasi parameter kemampuan untuk menghasilkan estimasi yang stabil dan pengaruh pembobotan butir RP dan butir RB terhadap kesalahan baku.

Penelitian dalam dua tahap, simulasi menggunakan 30 kondisi dengan replikasi 50 dengan variabel panjang tes, jumlah kategori, dan jumlah parameter butir, dan analisis deskriptif, dilanjutkan penerapan penskalaan gabungan butir tipe respons pilihan (rp) dan butir respons bebas (rb) pada konstruksi tes elektronika yang terdiri 40 butir pilihan ganda dan 4 butir jawaban tersusun, 3 butir memiliki lima kategori jawaban dan 1 butir dengan 4 kategori jawaban, melibatkan 355 siswa.

Hasil penelitian menunjukkan: ukuran sampel kurang berpengaruh pada *root mean square error* atau *(RSME)*, dan korelasi antara θ dengan $\hat{\theta}$, namun berpengaruh terhadap akurasi estimasi parameter butir pilihan ganda (b_{RP}) dan parameter butir respons tersusun (δ_{RP}). Jumlah parameter butir berpengaruh terhadap parameter kemampuan, tetapi tidak berpengaruh terhadap akurasi dari b_{RP} dan δ_{RP}. Estimasi dari parameter tingkat kesulitan butir jawaban tersusun tiga kategori lebih akurat daripada butir jawaban tersusun lima kategori. Estimasi tahan (*robust*) untuk parameter kesulitan butir jawaban tersusun 5 kategori memerlukan sampel minimal 250 responden, sedangkan untuk butir respons tersusun 3 kategori memerlukan sampel minimal 100 responden. Estimasi parameter kemampuan dari skor total (θ_{MCCR}) tidak sama dengan rata-rata jumlah *theta* dari masing-masing subtes $(\theta_{MC} + \theta_{CR})$. Theta dari tes yang dikalibrasi bersama-sama berbeda dengan *theta* dari total subtes yang dikalibrasi secara terpisah. Korelasi kemampuan yang mengunakan pembobotan dan kemampuan tanpa pembobotan mempunyai suatu rentang dari 0,988 sampai 0,948.

Kata kunci: *penyekalaan, model rash dan partial credit.*

Pendahuluan

Penilaian hasil belajar dapat dilakukan dengan berbagai format tes, antara lain tes respons butir dan uraian. Penilaian menggunakan butir tes respons pilihan, siswa memilih respons dari beberapa pilihan yang telah disediakan. Jenis tes respons pilihan meliputi pilihan ganda, benar-salah, maupun menjodohkan. Penggunaan tes bentuk respons pilihan memiliki keterbatasan, karena tidak semua kemampuan dapat diukur oleh butir format tersebut (Mardapi, 2001). Kemampuan yang memerlukan kinerja kompleks sukar diukur oleh butir tes respons pilihan. Ketika suatu kompetensi diukur oleh tes yang hanya terdiri atas butir respons pilihan, dapat terjadi sebagian konstruk tidak dapat diukur atau *construct underrepresentation* (Messick, 1995; AERA, APA & NCME, 1999). Penilaian yang hanya terdiri atas butir-butir respons pilihan sering dikritik karena kurang mencerminkan proses pembelajaran yang sebenarnya (Martinez, 1999; Simkin & Kuechler, 2005).

Suatu butir respons bebas mengacu pada format butir yang menghendaki peserta tes membuat atau menyusun sebuah respons bukan memilih jawaban yang sudah disediakan. Umumnya tes format respons bebas menghendaki siswa memberi respons satu atau lebih kalimat, merancang, dan melakukan penyelidikan maupun menjelaskan penyelesaian masalah yang memerlukan beberapa langkah. Jenis penilaian respons bebas meliputi jawaban singkat, uraian, tes kinerja, proyek, dan portofolio (Stecher *et al.*, 1997:25).

Program pengujian yang menggabungkan format respons pilihan dan format respons bebas (*polytomous*) dapat memperoleh manfaat yang bersifat saling melengkapi. Pada pengujian yang mengandung format respons pilihan, terdapat lebih banyak butir pertanyaan dalam suatu periode waktu. Dengan demikian, butir respons pilihan efisien dalam hal waktu pengujian (Wainer & Thissen, 1993). Program pengujian yang menggunakan format respons bebas berpotensi memperoleh informasi yang lebih dalam tentang kompetensi siswa. Demikian juga, butir respons bebas dapat memberikan format yang lebih sesuai untuk keterampilan tertentu, seperti pemecahan masalah. Penggabungan format respons pilihan dan respons bebas

memungkinkan terungkapnya kompetensi siswa secara luas dan mendalam. Pengujian gabungan dimaksudkan untuk menggabungkan kelebihan masing-masing tipe seperti penskoran yang obyektif, reliabilitas yang tinggi dari butir respons pilihan, dan validitas yang lebih baik dari butir respons bebas (Rudner, 2001).

Dalam banyak situasi pengujian, siswa sering diberi tes majemuk yaitu tes yang terdiri atas berbagai macam format dan hasilnya digabungkan untuk membentuk skor komposit tunggal. Pada penelitian ini, format tes yang digunakan adalah respons pilihan dan respons bebas. Penskalaan butir format respons pilihan dan respons bebas adalah menempatkan skor butir respons pilihan (RP) dan skor respons bebas (RB) pada skala yang sama (Ferrara, 1993). Menskala dua tipe butir tes tersebut menimbulkan tiga masalah utama yang perlu dikaji. Pertama, pengaruh ukuran sampel, panjang tes dan jumlah kategori respons terhadap akurasi estimasi parameter kemampuan (θ) dan parameter butir. Kedua, cara mengestimasi dua tipe butir tersebut, apakah butir RP dan RB diestimasi secara bersama-sama ataukah diestimasi secara terpisah. Ketiga, cara membobot relatif masing-masing tipe butir yang menghasilkan skor komposit.

Penelitian ingin menjawab beberapa pertanyaan berikut:

1. Apakah ukuran sampel mempengaruhi akurasi estimasi parameter kemampuan dan parameter butir?
2. Apakah jumlah parameter butir respons pilihan dan respons bebas mempengaruhi akurasi estimasi parameter kemampuan dan parameter butir?
3. Apakah jumlah kategori respons dari butir respons bebas mempengaruhi akurasi estimasi parameter butir?
4. Apakah terdapat perbedaan antara estimasi parameter kemampuan yang dihasilkan dari tes yang digabung serta dikalibrasi bersama-sama dan estimasi parameter kemampuan yang dihasilkan dari butir respons pilihan maupun respons bebas yang dikalibrasi secara terpisah?
5. Bagaimanakah hubungan antara kemampuan (θ) yang diestimasi dengan menggunakan pembobotan dan kemampuan (θ) yang diestimasi tanpa menggunakan pembobotan?

Untuk menjawab rumusan masalah nomor 1 sampai dengan 3, dilakukan penelitian simulasi Monte Carlo. Metode simulai merupakan cara terbaik menjawab pertanyaan tentang akurasi estimasi parameter kemampuan θ dan parameter butir. Untuk menjawab rumusan masalah nomor 4 dan 5, dilakukan penelitian empiris. Data yang diperoleh dari penelitian empiris dapat diperiksa kesesuaiannya dengan model. Data yang dianalisis adalah data yang sesuai (*fit*) dengan model.

Model *Rasch* dan *Partial Credit*

1. Model *Rasch*

Pada mulanya model teori respons butir dikembangkan untuk menangani butir yang diskor benar salah atau dikotomi. Pada teori respons butir, model matematika untuk kurva karakteristik butir (*item characteristic curve* = ICC) adalah bentuk kumulatif dan fungsi logistik. Terdapat tiga model yaitu model logistik satu parameter, dua parameter dan tiga parameter (1-PL, 2-PL, dan 3-PL). Dalam penelitian ini hanya difokuskan pada model satu parameter.

Rasch (1980:75) adalah orang pertama yang mengembangkan model logistik satu parameter. Pada model Rasch, orang diberi karakteristik tingkat kemampuan laten ξ dan butir diberi karakteristik tingkat kesukaran δ. Probabilitas menjawab benar suatu butir adalah fungsi dari perbandingan antara tingkat kemampuan dan kesukaran butir yang ditulis dengan simbol ξ/δ.

Untuk suatu butir tertentu, Rasch mengusulkan fungsi probabilitas sederhana sebagai

$$p = \frac{\xi}{\xi+\delta} \qquad (1)$$

Model dalam bentuk persamaan tersebut mempunyai tafsiran probabilitas jawaban benar (p) sama dengan nilai parameter kemampuan orang ξ relatif terhadap nilai parameter tingkat kesukaran butir δ (van der Linden & Hambleton, 1997:10). Jika digunakan notasi teori respons butir yang

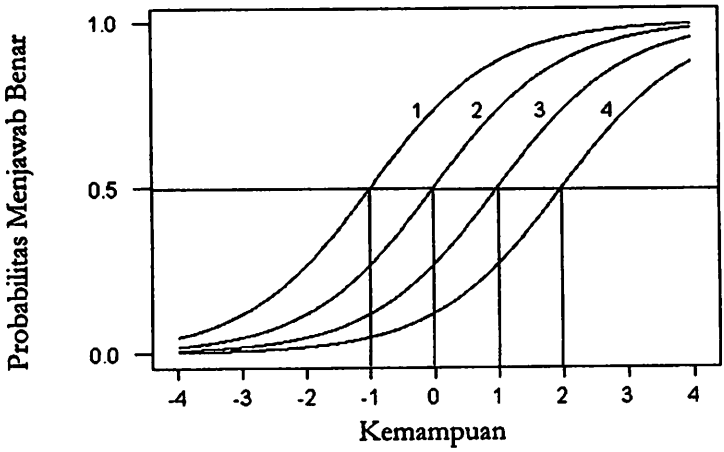
berlaku sekarang, ξ diganti dengan exp θ dan δ diganti dengan exp β, sehingga persamaan (1) berubah menjadi:

$$p(\theta) = \frac{\exp(\theta - \beta)}{1 + \exp(\theta - \beta)} \qquad (2)$$

Persamaan (2) menyatakan probabilitas menjawab benar dari responden yang mempunyai kemampuan θ menjawab suatu butir yang mempunyai tingkat kesukaran β (Hambleton *et al.*, 1991:13; Embretson & Reise, 2000:67): Parameter β adalah titik pada skala kemampuan dimana probabilitas menjawab benar adalah 0,5. Parameter tersebut merupakan parameter lokasi, yang menunjukkan posisi kurva karakteristik butir dalam hubungannya dengan skala kemampuan. Semakin besar nilai parameter β, semakin besar kemampuan yang diperlukan peserta tes untuk memperoleh 50% kemungkinan menjawab butir secara benar. Parameter β disebut juga parameter kesukaran butir.

Ketika nilai kemampuan dari sekelompok peserta tes ditransformasi sehingga reratanya adalah 0 dan simpangan bakunya adalah 1, maka nilai β berubah dari sekitar –2,0 sampai +2,0. Nilai β yang mendekati –2,0 berarti butir yang sangat mudah, dan nilai β yang mendekati +2,0 berarti butir yang sangat sukar. Beberapa contoh kurva karakteristik butir model satu parameter diperlihatkan pada Gambar 1. Parameter butir kurva tersebut adalah sebagai berikut: untuk butir 1, β_1 =-1,0; untuk butir 2, β_2 = 0,0; untuk butir 3, β_3 = 1,0; dan untuk butir 4, β_4 = 2,0.

Ciri penting model Rasch adalah model tersebut tidak mengandung parameter diskriminasi. Rumus Rasch menghitung probabilitas logistik dengan kemiringan yang sama. Kurva-kurva berbeda hanya pada lokasinya pada skala kemampuan. Pada model Rasch, diasumsikan bahwa kesukaran butir merupakan satu-satunya karakteristik butir yang mempengaruhi kinerja peserta tes.



Gambar 1. Kurva Karakteristik Butir Model Rasch (Hambleton *et al.*, 1991:14).

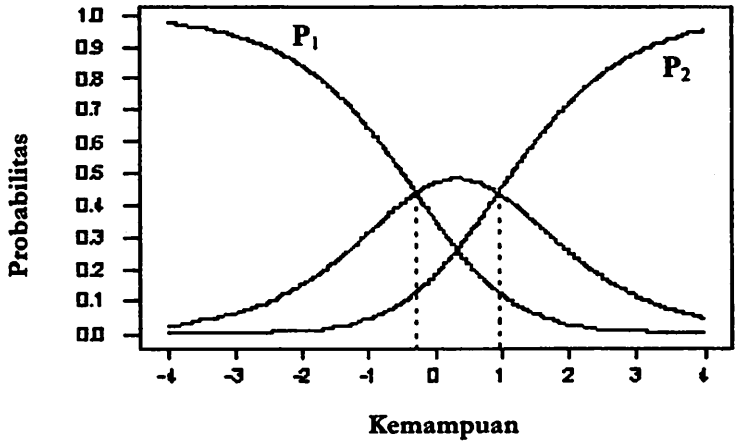
2. Model *Partial Credit*

Beberapa model telah dikembangkan untuk menangani butir tes yang diskor secara *polytomous*, diantaranya adalah model respons berjenjang atau *graded response* (Samejima, 1997:85), model respons nominal atau *nominal response* (Bock, 1997:33), model kredit parsial atau *partial credit* (PC). Model PC mempunyai sifat yang sama dengan model Rasch yaitu daya beda tiap butir tes dianggap sama.

Model PC (Masters, 1999:101) dikembangkan untuk menganalisis butir-butir tes yang memerlukan banyak langkah dalam proses penyelesaiannya. Model tersebut dapat juga digunakan untuk menganalisis respons skala sikap (Embretson & Reise, 2000:105). Model PC dapat dianggap sebagai perluasan model Rasch. Butir i diskor $x = 0, 1, \ldots, m_i$ untuk suatu butir dengan k kategori respons ($k = m_i + 1$). Masters (1982, 1999:101) mengusulkan persamaan umum tentang probabilitas peserta tes yang mempunyai kemampuan θ dan memperoleh skor x pada butir i sebagai berikut:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^{x}(\theta - \delta_{ij})\right]}{\sum_{k=0}^{m_i}\left[\exp\sum_{j=0}^{k}(\theta - \delta_{ij})\right]} \qquad x = 0, 1, \ldots, m_i \qquad (3)$$

Notasi δ_{ij} disebut kesukaran langkah (*step difficulty*) butir yang berkaitan dengan transisi dari satu kategori (skor) ke kategori berikutnya. Untuk konvensi notasi, $\sum (\theta - \delta_{ij})$ ditentukan sama dengan nol ketika $k = 0$ (Embretson & Reise, 2000:105). Gambar 2 memperlihatkan kurva karakteristik operasi (*operating characteristic curve*) butir tes yang mempunyai tiga kategori jawaban dengan $\delta_1 = -0{,}29$ dan $\delta_2 = 0{,}95$.



Gambar 2. Kurva Respons Kategori pada Model *Partial Credit* (Embretson & Reise, 2000:107)

Metode Penelitian

Penelitian dilakukan dua tahap. Tahap pertama penelitian dilakukan melalui simulasi. Penelitian pada tahap tersebut dilakukan untuk (1) menyelidiki pengaruh jumlah parameter butir terhadap akurasi estimasi parameter, (2) menyelidiki pengaruh jumlah kategori respons dari butir

respons bebas terhadap akurasi estimasi parameter, (3) menyelidiki pengaruh ukuran sampel terhadap akurasi estimasi parameter kemampuan dan parameter butir, dan (4) menyelidiki ukuran sampel yang akan menghasilkan estimasi yang stabil untuk butir respons bebas (RB) dan respons pilihan (RP).

Penelitian tahap kedua merupakan penerapan penskalaan gabungan butir tipe respons pilihan (RP) dan butir respons bebas (RB) pada konstruksi tes elektronika. Penelitian pada tahap tersebut dilakukan untuk (1) menyelidiki estimasi parameter kemampuan yang dihasilkan dari tes yang digabung dan dikalibrasi bersama-sama dan estimasi parameter kemampuan dari butir RP maupun RB yang dikalibrasi secara terpisah, dan (2) menyelidiki pengaruh pembobotan butir RP dan butir RB terhadap kesalahan baku.

Analisis data, yaitu: (1) melakukan analisis subtes yang dikalibrasi bersama dan terpisah, (2) melakukan analisis skor tes terbobot dan tidak terbobot, dan (3) menghitung fungsi informasi dan kesalahan baku pengukuran.

Hasil Penelitian

1. Hasil Penelitian Simulasi

a. Pembuatan Data Dikotomi

Untuk membangkitkan data respons peserta tes perlu ditentukan parameter butir. Parameter butir dibangkitkan menurut distribusi *uniform*. Parameter tersebut digunakan sebagai pedoman untuk pembangkitan data respons. Pembangkitan data dikelompokkan menjadi dua jenis: dikotomi atau respons pilihan (RP) dan *polytomous* atau respons bebas (RB). Parameter kesukaran tes tipe RP yang dibangkitkan terdiri dari 60 buah dengan rentang kesukaran butir (b) dari –2 sampai +2.

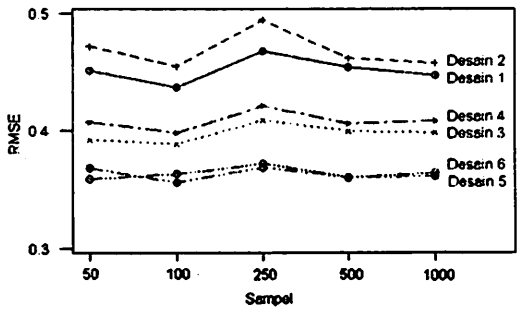
b. Pembuatan Data *Polytomous*

Parameter butir tes tipe respons bebas (RB) terdiri dari dua kumpulan. Kumpulan pertama adalah 10 buah butir 3 level kategori dengan rentang kesukaran δ_1 dari –2 sampai 0, δ_2 dari 0 sampai +2.

Kumpulan ketiga adalah 5 buah butir 5 level kategori dengan rentang kesukaran δ_1 dari –2 sampai –1, δ_2 dari –1 sampai 0, δ_3 dari 0 sampai +1, dan δ_4 dari +1 sampai +2.

c. RMSE Parameter Butir dan Parameter Kemampuan

Untuk mengevaluasi akurasi estimasi digunakan akar dari dari rerata kesalahan kuadrat (*root mean squared error* = RMSE). Parameter yang dibandingkan adalah parameter sejati dan parameter estimasi. Ada dua macam parameter sejati maupun parameter estimasi yaitu parameter kesukaran butir dan parameteter kemampuan (θ). Parameter butir sejati dibangkitkan menurut distribusi *uniform*. Sedangkan parameter kemampuan sejati merupakan bilangan acak dari distribusi normal dengan rerata nol dan simpangan baku satu. Jumlah parameter butir bervariasi bergantung jenis desainnya. Parameter butir estimasi berasal dari program QUEST yang mengestimasi data respons hasil pembangkitan dengan basis parameter sejati. Simulasi dilakukan sebanyak 50 replikasi. Gambar 3 memperlihatkan rangkuman perhitungan RMSE berbagai desain.

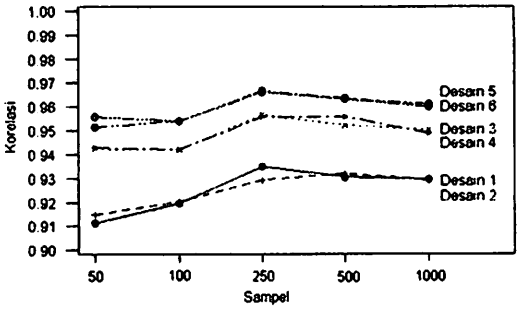


Gambar 3. Grafik Rerata RMSE Parameter Kemampuan (Theta) untuk Berbagai Desain.

d. Korelasi Parameter Butir dan Parameter Kemampuan

Cara lain untuk mengevaluasi akurasi estimasi adalah korelasi *product moment*. Dilakukan analisis korelasi antara parameter butir sejati dan

parameter butir estimasi, serta parameter kemampuan sejati dan parameter kemampuan estimasi. Gambar 4 memperlihatkan rangkuman perhitungan RMSE berbagai desain.



Gambar 4. Grafik Rerata Korelasi Parameter Kemampuan (θ) untuk Berbagai Desain.

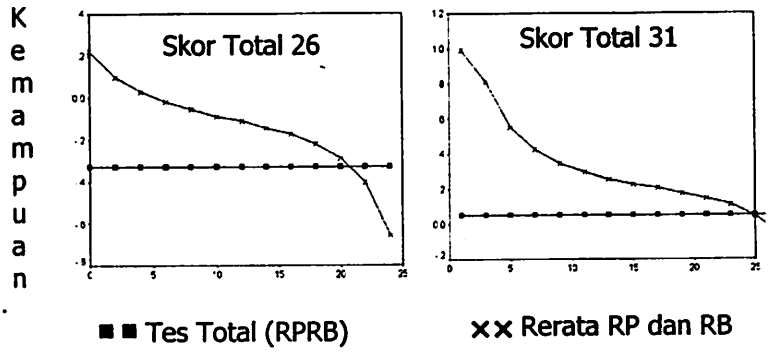
2. Hasil Penelitian Empiris

a. Estimasi Kemampuan dari Subtes yang Dikalibrasi Bersama dan Terpisah

Tes yang terdiri atas butir-butir respons pilihan (RP) disebut subtes RP dan tes yang terdiri atas butir-butir respons bebas (RB) disebut subtes RB. Ada dua cara mengalibrasi tes elektronika. Pertama, butir-butir RP dan RB dikalibrasi bersama-sama sehingga diperoleh parameter kemampuan θ_{RPRB}. Kedua, masing-masing subtes RP dan subtes RB dikalibrasi secara terpisah sehingga diperoleh θ_{RP} dan θ_{RB}.

Setiap skor subtes menyatakan ukuran kemampuan orang demikian juga skor tes total. Ukuran (*measure*) yang digunakan di lingkungan pengukuran model Rasch merupakan transformasi skor mentah yang mempunyai satuan logit (Wright & Linacre, 1989; Wright & Douglas, 1996; Linacre, 1998). Jika setiap orang menjawab secara konsisten, ukuran kemampuan dari skor setiap subtes akan sama dengan ukuran kemampuan

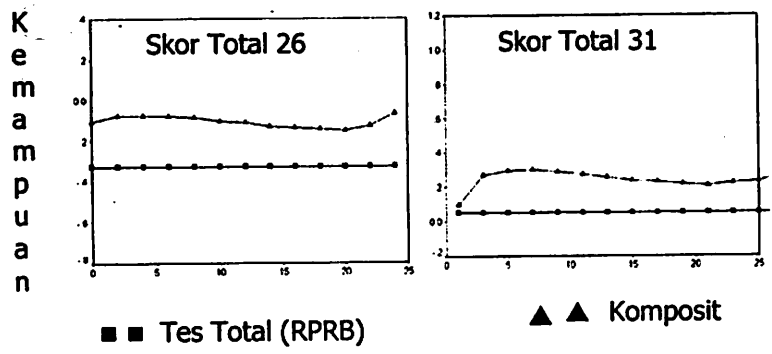
dari skor tes total. Dengan kata lain rerata dari jumlah ukuran kemampuan skor subtes akan sama dengan ukuran kemampuan dari skor tes total. Kenyataannya tidak selalu terjadi demikian. Gambar 5 menjelaskan bahwa rerata dari jumlah ukuran kemampuan skor subtes tidak sama dengan ukuran kemampuan dari skor tes total



Gambar 5. Ukuran Subtes dan Tes Total

Ketika pengguna tes ingin mengetahui skor setiap subtes, maka tes harus dikalibrasi secara terpisah. Rerata dari jumlah θ_{RP} dan θ_{RB} untuk skor total 26 adalah –0,02; -0,05 dan –0,09. Skor total 26 mempunyai nilai konversi θ yang berbeda-beda jika subtes tipe RP dan RP dikalibrasi secara terpisah. Gambar 9 memperlihatkan rerata dari θ_{RP} + θ_{RB} yang bervariasi. Setiap perbedaan antara skor RP dan RB mempunyai nilai rerata yang berlainan.

Aproksimasi untuk memperoleh nilai θ komposit (θ_{RPRB}) dari θ subtes (θ_{RP} dan θ_{RB}) dapat dihitung dengan menggunakan invers dari rerata kesalahan baku setiap θ subtes (Wright, 1994). Gambar 6 memerlihatkan grafik dari skor komposit tersebut.



Gambar 6. Ukuran Komposit dan Tes Total

Satu skor total RPRB dapat terdiri atas berbagai kombinasi skor RP dan RB. Ketika pengguna tes menginginkan nilai akhir maka skor RP dan skor RB dapat dikalibrasi secara bersama-sama. Ketika pengguna tes menginginkan nilai masing-masing subtes, maka skor RP dan skor RB dapat dikalibrasi secara terpisah. Jika pengguna tes menginginkan nilai masing-masing subtes dan nilai akhir, maka nilai akhir dapat diperoleh dengan cara menghitung nilai komposit dengan menggunakan invers kesalahan baku.

b. Pembobotan

Pada analisis menggunakan pembobotan, skor butir format RP dan RB dianalisis dan dilaporkan bersama-sama, yang menghasilkan satu skor komposit. Rasio pembobotan dicantumkan pada tabel 1.

Tabel 1
Rasio Pembobotan

Rasio Pembobotan	Bobot Tipe Respons Pilihan (RP)	Bobot Tipe Respons Bebas (RB)
4:1	1,100	0,733
3:1	1,031	0,917
2:1	0,916	1,222
1:1	0,688	1,833
1:2	0,458	2,442
1:3	0,344	2,750
1:4	0,275	2,933

Setelah dihitung nilai pembobotan setiap rasio, selanjutnya dihitung nilai prediksi skor RP dan RB untuk sesetiap pembobotan. Nilai prediksi tersebut dianggap sebagai skor komposit. Nilai komposit terbobot selanjutnya dikonversi ke skor terskala. Transformasi yang digunakan adalah transformasi linier. Statistik deskriptif skor terskala dirangkum pada tabel 2.

Tabel 2
Statistik Skor Terskala

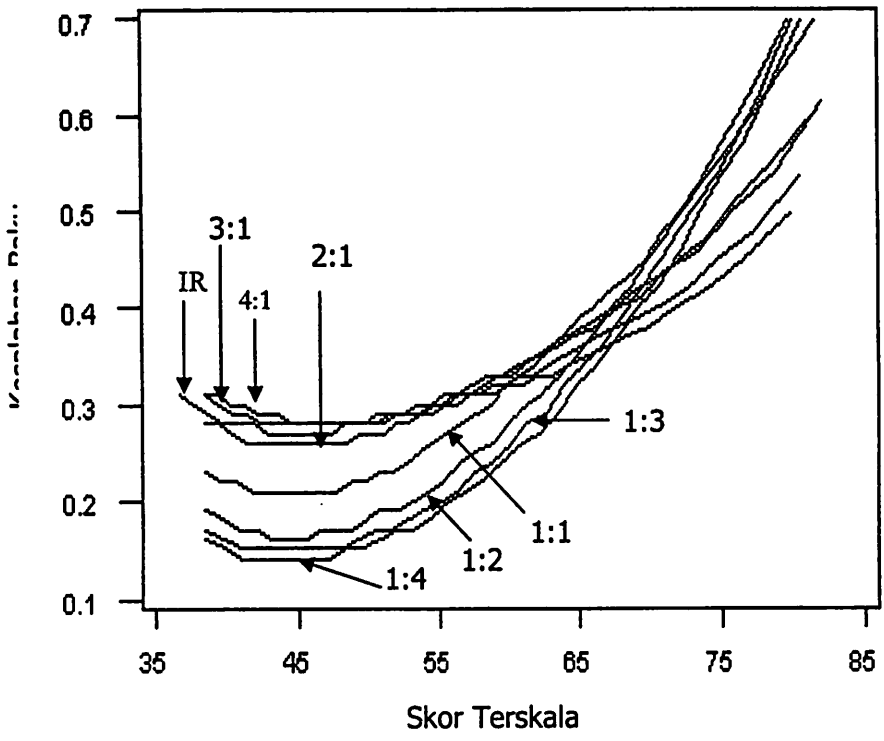
Model	Rerata	Simpangan Baku	Minimum	Maksimum
IRT	48.29	8.19	36.76	82.25
Mentah	49.58	15.08	25.45	94.55
Pem 4:1	48.37	14.80	25.42	94.11
Pem 3:1	49.20	14.99	25.47	94.42
Pem 2:1	50.54	15.32	25.56	94.85
Pem 1:1	53.34	16.03	25.78	95.95
Pem 1:2	56.42	17.95	25.24	97.51
Pem 1:3	57.48	17.16	26.07	97.49
Pem 1:4	58.30	17.39	26.13	97.78

Skor ekspektasi yang sudah diskala dikorelasikan dengan skor mentah dan skor IRT (θ) yang sudah diskala. Hasil korelasi ditunjukkan pada tabel 4. Pembobotan dengan rasio 4:1, 3:1 dan 2:1 mempunyai korelasi mendekati 1,000 dengan skor IRT maupun skor mentah.

Kurva kesalahan baku untuk sesetiap model penskoran diperlihatkan pada gambar 7. Kesalahan baku merupakan kebalikan dari akar kuadrat fungsi informasi tes. Sebelum menghitung fungsi informasi tes, skor prediksi terbobot dikonversi ke satuan logit.

Tabel 3
Rerata Kesalahan Baku Korelasi Skor Terbobot dengan Skor IRT, Skor Mentah dan Skor Mentah Terbobot

Model	Rerata Kesalahan Baku	Korelasi		
		Skor IRT	Skor Mentah	Skor Mentah Terbobot
Pem 4:1	0.296	0,988	1,000	0,996
Pem 3:1	0.295	0,986	1,000	1,000
Pem 2:1	0.287	0,981	1,000	0,997
Pem 1:1	0.265	0,970	0,997	0,967
Pem 1:2	0.248	0,952	0,991	0,917
Pem 1:3	0.246	0,952	0,990	0,888
Pem 1:4	0.245	0,948	0,988	0,871



Gambar 7. Kesalahan Baku Berbagai Model Penskoran

Pembahasan

Program Quest (Adam & Khoo, 1996) menggunakan metode JMLE (*joint maximum likehood estimation*) untuk mengestimasi parameter kemampuan dan parameter butir. Ketika parameter kemampuan dan butir diestimasi secara serentak, tidak ada bukti matematis pada ukuran sampel berapakah estimasi parameter menjadi akurat (Hulin *et al.*, 1983:99). Cara mengetahui akurasi estimasi adalah melalui simulasi. Faktor-faktor yang diteliti pada simulasi adalah jumlah parameter yang terdiri atas 40, 56 dan

72 parameter butir, jumlah kategori yang terdiri atas 3 kategori dan 5 kategori, dan ukuran sampel yang terdiri atas 50, 100, 250, 500 dan 1000 sampel.

Jumlah parameter butir merupakan fungsi dari panjang tes. Jumlah parameter butir 40 dapat berasal dari gabungan 20 butir RP dan 10 butir RB 3 kategori atau berasal dari gabungan 20 butir RP dan 5 butir RB 5 kategori. Pada parameter 40 dengan panjang tes 20 RP dan 10 RB 3 kategori, RMSE θ adalah 0,45 untuk ukuran sampel 50 dan 0,44 untuk ukuran sampel 1000; korelasi θ adalah 0,91 untuk ukuran sampel 50 dan 0,92 untuk ukuran sampel 1000. Pada parameter 40 dengan panjang tes 20 RP dan 5 RB 5 kategori, RMSE θ adalah 0,47 untuk ukuran sampel 50 dan 0,45 untuk ukuran sampel 1000; korelasi θ adalah 0,91 untuk ukuran sampel 50 dan 0,92 untuk ukuran sampel 1000.

Jumlah parameter butir 72 dapat berasal dari gabungan 60 butir RP dan 6 butir RB 3 kategori atau berasal dari gabungan 60 butir RP dan 3 butir RB 5 kategori. Pada parameter 72 dengan panjang tes 60 RP dan 6 RB 3 kategori, RMSE θ adalah 0,36 untuk ukuran sampel 50 dan 1000; korelasi θ adalah 0,95 untuk ukuran sampel 50 dan 0,96 untuk ukuran sampel 1000. Pada parameter 72 dengan panjang tes 60 RP dan 3 RB 5 kategori, RMSE θ adalah 0,35 untuk ukuran sampel 50 dan 0,36 untuk ukuran sampel 1000; korelasi θ adalah 0,95 untuk ukuran sampel 50 dan 0,96 untuk ukuran sampel 1000. Untuk parameter θ, ukuran sampel bukan merupakan faktor utama. Menaikkan ukuran sampel tidak akan menurunkan RMSE maupun menaikkan korelasi θ dan $\hat{\theta}$. Menurunkan ukuran sampel tidak akan menaikkan RMSE maupun menurunkan korelasi θ dan $\hat{\theta}$. Pada sampel 50, estimasi parameter kemampuan sama akuratnya seperti pada sampel 1000.

Parameter kesukaran butir dibedakan menjadi dua: kesukaran butir RP, b_{RP} dan kesukaran langkah butir RB, δ_{RB}. Pada parameter b_{RP}, ketika ukuran sampel berubah dari 1000 menjadi 50, nilai RMSE berubah naik dua kali. Ukuran sampel mempengaruhi 95% variabilitas RMSE b_{RP} dan 97% variabilitas korelasi b_{RP}. Jumlah parameter butir tidak mempengaruhi

akurasi estimasi b_{RP}. Pada parameter δ_{RB}, nilai RMSE bertambah lebih dari dua kali ketika sampel berubah dari 1000 menjadi 50. Ukuran sampel mempengaruhi 68% variabilitas RMSE δ_{RB} dan 30% variabilitas korelasi δ_{RB}. Jumlah parameter butir tidak mempengaruhi akurasi estimasi parameter δ_{RB}.

Parameter kesukaran butir yang dikenal sebagai parameter struktural (Swaminathan, 1983: 33) tidak harus konsisten, kenaikan jumlah parameter butir tidak menaikan akurasi estimasi parameter butir. Parameter kemampuan yang dikenal sebagai parameter insidental bersifat konsisten. Menaikkan ukuran sampel akan menaikkan akurasi estimasi parameter butir. Jumlah parameter butir yang tidak mempengaruhi akurasi estimasi parameter butir sesuai dengan penelitian DeMars (2003) yang menggunakan program MULTILOG untuk mengestimasi parameter butir model nominal.

Ada perbedaan signifikan nilai RMSE dan korelasi b_{RB} yang disebabkan oleh jumlah kategori. Estimasi parameter δ_{RB} dari tes butir respons bebas 3 kategori lebih akurat daripada estimasi parameter δ_{RB} dari tes butir respons bebas 5 kategori. Estimasi kesukaran langkah terakhir dari butir respons bebas 5 kategori (δ_4) mempunyai akurasi yang jelek untuk ukuran sampel 50 dan 100. Jumlah kategori yang semakin banyak memerlukan responden yang semakin besar pula. Untuk memperoleh stabilitas yang memadai diperlukan sedikitnya 25 * (jumlah kategori) responden (Linacre, 2002a). Sampel 250 dapat menghasilkan parameter kesukaran langkah terakhir (δ_4) cukup akurat.

Subtes respons pilihan terdiri atas 40 butir yang diskor 0 dan 1, sehingga skor total adalah 40 poin. Subtes respons bebas terdiri atas 4 butir, 3 butir diskor maksimum 4 poin, 1 butir diskor maksimum 3 poin, sehingga skor subtes respons bebas adalah 15 poin. Skor kedua subtes tersebut merupakan data ordinal. Analisis Rasch mentransformasi skor ordinal tersebut ke skala logit sehingga menjadi data interval. Karena sebagian besar teknik statistik mengasumsikan bahwa data harus bersifat interval, skor transformasi Rasch dapat digunakan untuk berbagai keperluan.

Tes total terdiri atas subtes respons pilihan dan subtes respons bebas. Setiap skor subtes menyatakan kemampuan seseorang seperti juga skor tes total. Jika peserta menjawab tes secara konsisten, ukuran kemampuan (θ) dari skor setiap subtes akan sama dengan θ dari skor tes total. Estimasi kemampuan yang diperoleh dari skor total (θ_{RPRB}) tidak selalu sama dengan rerata dari jumlah theta setiap subtes ($\theta_{RP}+\theta_{RB}$). Nilai rerata ($\theta_{RP}+\theta_{RB}$) selalu bervariasi meskipun jumlah skor subtes RP dan subtes RB tetap. Theta (θ) yang dihasilkan dari tes yang dikalibrasi bersama-sama berbeda dengan theta yang dihasilkan dari jumlah subtes yang dikalibrasi secara terpisah.

Perbedaan antara θ_{RPRB} dan rerata ($\theta_{RP}+\theta_{RB}$), disebabkan pengaruh ketidak simetrisan transformasi skor mentah menjadi nilai θ. Ketika rasio jawaban benar-salah (B/S) subtes RP dan RB sama dengan rasio jawaban benar-salah tes total, maka rerata ($\theta_{RP}+\theta_{RB}$) sama dengan θ_{RPRB}. Ketika skor dua subtes berbeda jauh, meskipun skor total tetap sama, rerata ($\theta_{RP}+\theta_{RB}$) menjadi besar. Rerata menjadi takterhingga ketika salah satu skor subtes adalah sempurna (salah semua maupun betul semua). Pengaruh ketidak-simetrisan tersebut akan meningkatkan varian dari θ subtes, yang menyebabkan distribusi rerata ($\theta_{RP}+\theta_{RB}$) lebih besar daripada distribusi θ_{RPRB}. Nilai θ_{RPRB} mempunyai rentang 4,33 dan varian 0,61. Sedangkan rerata ($\theta_{RP}+\theta_{RB}$) mempunyai rentang 4,75 dan varian 0,68. Untuk mengurangi pengaruh ketidak-simetrisan, nilai θ_{RPRB} dihitung dengan memasukan invers dari rerata kesalahan baku. Pengaruh ketidak-simetrisan subtes diteliti juga oleh Bowles (1999). Ada tiga subtes yang masing-masing mempunyai tingkat kesukaran tinggi, sedang dan rendah. Skor subtes sedang dan rendah tidak simetris. Dengan menggunakan penskoran yang memasukkan faktor invers rerata kesalahan baku, diperoleh skor komposit yang konstan.

Terdapat tujuh variasi pembobotan, nilai rasio pembobotan 4:1 sampai 1:4. Pada pembobotan rasio 4:1, bobot butir RP adalah 4/5 kali (skor mak gabungan/skor mak RP) diperoleh nilai 1,100 dan bobot butir

RB adalah 1/5 kali (skor mak gabungan/skor mak RB) diperoleh nilai 0,733.

Skor tes RP yang dibobot mempunyai nilai maksimum 44, sedangkan skor tes RB mempunyai nilai maksimum 11. Pada pembobotan rasio 1:4, bobot butir RP adalah 1/5 kali (skor mak gabungan/skor mak RP) diperoleh nilai 0,275 dan bobot butir RB adalah 4/5 kali (skor mak gabungan/skor mak RB) diperoleh nilai 2,933. Skor tes RP yang dibobot mempunyai nilai maksimum 11, sedangkan skor tes RB mempunyai nilai maksimum 44. Pada semua variasi pembobotan, skor RP dan RB yang sudah dibobot memberi sumbangan ke skor total yang sama yaitu 55. Pembobotan akan mempengaruhi fungsi informasi mapun kesalahan baku pengukuran. Fungsi informasi berbanding langsung dengan kuadrat pembobotan. Kesalahan baku pengukuran berbanding terbalik dengan kuadrat pembobotan. Pada pembobotan rasio 4:1, butir RP dibobot lebih besar daripada butir RB. Pada pembobotan rasio 1:4, butir RB dibobot lebih besar daripada butir RP. Rentang pembobotan dari rasio 4:1 sampai rasio 1:4 menyatakan pembobotan butir RP berubah dari besar menjadi kecil dan pembobotan butir RB berubah dari kecil menjadi besar.

Bobot butir RB lebih berpengaruh terhadap fungsi informasi maupun kesalahan baku pengukuran. Fungsi informasi dengan pembobotan rasio 1:4 lebih besar daripada fungsi informasi dengan pembobotan rasio 4:1. Kesalahan baku pengukuran dengan pembobotan rasio 1:4 lebih kecil daripada kesalahan baku pengukuran dengan pembobotan rasio 4:1. Pembobotan rasio 4:1 mengakibatkan rerata kesalahan baku pengukuran sebesar 0,296. Pembobotan rasio 1:4 mempunyai rerata kesalahan baku 0,245. Semakin besar bobot pada butir RB semakin kecil kesalahan baku pengukuran.

Korelasi skor IRT dan skor berdasarkan pembobotan mempunyai rentang dari 0,988 sampai 0,948. Nilai korelasi mendekati satu berarti skor IRT dan skor berdasarkan pembobotan tidak mempengaruhi urutan peringkat peserta tes. Peserta urutan pertama berdasarkan skor IRT akan tetap berada pada urutan tersebut meskipun didasarkan pada skor pembobotan. Hasil penelitian tersebut sesuai dengan hasil penelitian Childs *et al* (2004). Mereka membandingkan kesalahan baku pembobotan rasio

3:3:4 dan 2:2:6. Angka pertama, kedua dan ketiga dari rasio tersebut menyatakan proporsi bobot butir respons pilihan, respons bebas singkat dan respons bebas panjang. Kesalahan baku pembobotan rasio 2:2:6 lebih kecil daripada kesalahan baku pembobotan rasio 3:3:4. Schaeffer & Bene (2002) mengungkapkan bahwa tes yang dibobot mempunyai kesalahan baku lebih kecil daripada kesalahan baku tes yang tidak dibobot. Ito & Sykes (2000) menyatakan bahwa pembobotan dua kali pada butir tes respons bebas menurunkan kesalahan baku. Sykes & Hou (2003) menyatakan bahwa pembobotan butir tes respons pilihan menurunkan kesalahan baku pada skala bagian bawah.

Simpulan

Berdasarkan deskripsi data hasil penelitian dan pembahasan yang telah dijelaskan, dapat diambil kesimpulan sebagai berikut:

1. Ukuran sampel tidak mempengaruhi RMSE parameter kemampuan dan korelasi θ dan $\hat{\theta}$. Menaikkan ukuran sampel tidak akan menurunkan RMSE maupun menaikkan korelasi θ dan $\hat{\theta}$, demikian pula menurunkan ukuran sampel tidak akan menaikkan RMSE maupun menurunkan korelasi θ dan $\hat{\theta}$. Estimasi parameter kemampuan pada sampel 50 sama akuratnya seperti pada sampel 1000. Ukuran sampel mempengaruhi akurasi estimasi parameter kesukaran butir dikotomi (b_{RP}) dan *polytomous* (δ_{RB}). Nilai RMSE b_{RP} bertambah lebih dari dua kali ketika sampel berubah dari 1000 menjadi 50.
2. Jumlah parameter butir atau panjang tes mempengaruhi akurasi parameter θ, tetapi tidak mempengaruhi akurasi parameter b_{RP} dan δ_{RB}. RMSE θ turun ketika panjang tes atau jumlah parameter butir naik dan korelasi θ dan $\hat{\theta}$ naik ketika panjang tes atau jumlah parameter butir naik.
3. Estimasi parameter kesukaran langkah butir *polytomous* dari butir tes 3 kategori lebih akurat daripada estimasi parameter kesukaran langkah dari butir tes 5 kategori. Untuk memperoleh hasil estimasi parameter kesukaran butir yang stabil dari butir tes respons bebas 5 kategori

minimum diperlukan sampel 250 responden, dan dari butir tes respons bebas 3 kategori minimum diperlukan sampel 100 responden.

4. Estimasi kemampuan yang diperoleh dari skor total (θ_{RPRB}) tidak selalu sama dengan rerata dari jumlah θ setiap subtes ($\theta_{RP}+\theta_{RB}$). Nilai rerata ($\theta_{RP}+\theta_{RB}$) selalu bervariasi meskipun jumlah skor subtes RP dan subtes RB tetap. Theta yang dihasilkan dari tes yang dikalibrasi bersama-sama berbeda dengan θ yang dihasilkan dari jumlah subtes yang dikalibrasi secara terpisah.
5. Korelasi antara kemampuan yang diestimasi dengan pembobotan dan kemampuan yang diestimasi tanpa pembobotan mempunyai rentang dari 0,988 sampai 0,948. Pembobotan mempengaruhi fungsi informasi maupun kesalahan baku pengukuran. Semakin besar bobot pada butir RB, semakin kecil kesalahan baku pengukuran.

Daftar Pustaka

Adams, R., & Khoo, S.T. (1996). Quest: an Interactive Item Analysis Program. Melbourne: The Australian Council for Educational Research.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standard for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bock, R.D. (1997). The nominal categories model. Dalam W.J. van der Linden & Hambleton, R.K (Eds.), *Handbook of modern item response theory.* New York: Springer.

Bowles, R. (1999). Combining and dropping subtest measures. *Rasch Measurement Transactions, 13:1,* 686.

Childs, R.A., Elgie, S., Gadalla, T., & Traub, R. (2004). Irt-linked standard errors of weighted composites. *Practical Assessment, Research & Evaluation, 9*(13). Diambil pada tanggal 31 Oktober 2004, dari http://PAREonline.net/ getvn.asp?v=9&n=13.

DeMars, C.E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement, 27*, 275-288.

Embretson,S.E., & Reise,S.P. (2000). *Item response theory for psychologists.* Mahwah: Lawrence Erlbaum Associates.

Ferrara, S. (1993). *Generalizability theory and scaling: Their roles in writing assessment and implications for performace assessment in other content areas.* Paper presented at the annual meeting of the National Council on Measurement in Education, April 13, 1993, in Atlanta.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park: Sage Publications

Ito, K., & Sykes, R.C. (2000). *An evaluation of intentional weighting of extended-response or constructed-response items in tests with mixed item types.* Paper presented at the annual National Conference on Large Sxale Assessment, Snowbird, Utah.

Krathwohl, D.R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212-218.

Linacre, J.M. (1998). Estimating measures with known polytomous item difficulties. *Rasch Measurement Transaction, 12*, 638.

Linacre, J.M. (2002a). Optimizing rating scale category effectiveness. *Journal of Applied Measurement,* 3(1), 85-106.

Djemari Mardapi. (2000). *Perencanaan tes.* Makalah disampaikan pada pelatihan TOT Widyaiswara PPT Migas, Cepu, 1-12 Juli 1996.

Martinez, M.E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Masters, G.N. (1982). A Rasch-model for partial credit scoring. *Psychometrika, 47*, 149-174.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741-749.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Rudner, L.M. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice, 20*(1), 16-19.

Samejima, F. (1997). The graded response model. Dalam W.J. van der Linden & Hambleton, R.K (Eds.), *Handbook of modern item response theory.* New York: Springer.

Schaeffer, G.A., & Bene, N.H. (2002). A comparison of three scoring methods for test with selected-response and constructed-response items. *Educational Assessment, 8*(4), 317-340.

Simkin, M.G., & Kuechler, W.L. (2005). Multiple-choice tests and student understanding: What is the connection?. *Decission Sciences Journal of Innovative Education, 3*(1), 73-97.

Stecher,B.M., Rahn,M.L., Ruby,A., Alt,M.N., Robyn,A.E., & Ward,B. (1997). *Using alternatif assessments in vocational education.* Santa Monica: RAND.

Swaminathan, H. (1983). Parameter estimation in item response models. Dalam Ronald K. Hambleton (Ed.), *Applications of item response theory.* Vancouver, BC: Educational Research Institute of British Columbia.

Sykes, R.C., & Hou, L. (2003). Weighting constructed-response items in IRT-based exams. *Applied Measurement in Education, 16*(4), 257-275.

van der Linden, W. J., & Hambleton, R.K. (1997). *Handbook of modern item response theory.* New York: Springer.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*, 103-118.

Wright, B.D., & Douglas, G.A. (1996). Estimating measures with known item difficulties. *Rasch Measurement Transactions, 10*, 499.

Wright, B.D., & Linacre, J.M. (1989). Differences between scores and measures. *Rasch Measurement Transactions, 3*, 63.