PENGUKURAN UNJUK KERJA MENGGUNAKAN MODEL POLITOMUS

Emy Budiastuti
Jurusan PTBB FT UNY

ABSTRAK

Banyak instrumen pengukuran, khususnya instrumen untuk mengukur kinerja atau prestasi siswa pada mata pelajaran tertentu, memerlukan respon lebih dari dua kategori. Model politomus dari Item Response Theory (IRT), digunakan untuk menganalisis skor item yang memiliki respon multi kategori. Tes tersebut memberikan informasi lebih banyak dibandingkan dengan tes yang diskor secara dikotomus

Banyak model yang menggunakan pensekoran politomus, diantaranya adalah: the Graded Response Model (GRM), Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), RSM, dan NRMm (Embretson & Reise, 2000). Semua model tersebut mendasarkan pada asumsi bahwa respon (jawaban) peserta tes pada suatu item tes tergantung kemampuan peserta. Menggunakan simulasi data akan diperoleh estimasi kemampuan, kecenderungan, dan threshold.

Penilaian secara IRT atau secara politomus, memerlukan ketelitian dari rater atau penilai. Hasil penilaian yang akurat ditentukan kemampuan dan watak penilai atau *rater*. Dengan demikian diperlukan pemantauan berkesinambungan dalam proses menilai, pembuatan rubrik tergambar dengan jelas, dan pelatihan perlu ditingkatkan untuk rater.

Kata kunci : pengukuran unjuk kerja, model politomus

PENDAHULUAN

Dalam pengukuran pendidikan, dikenal teori pengukuran klasik dan teori pengukuran modern. Teori pengukuran klasik telah banyak berjasa dalam dunia pengukuran dan bahkan masih banyak digunakan sampai sekarang. Namun demikian dalam teori pengukuran klasik terdapat keterbatasan karena bersifat *group dependent* dan *item dependent* (Hambleton dan Swaminatan, 1991).

Item Response Theory (IRT) berdasarkan teori tentang perilaku latent, yang menggabungkan pengukuran terhadap examinee, baik tes maupun performans. Bagaimana performans dikaitkan dengan pengetahuan melalui item tes pada pengukuran. Melalui tata kerja IRT, dapat diformulasikan beberapa model pengukuran. Biasanya dikaitkan dengan model variasi scoring, seperti: dichotomous, binomial, Poisson, rating

scale, facets, multinomial logit, atau polytomous (Schumacker, 2005)

Model politomus dari IRT, digunakan untuk menganalisis skor item yang memiliki respon multi kategori. Contoh skor politomus adalah skala Likert untuk skala sikap atau juga respon tes essay. Dalam hal ini model partial credit kerap digunakan untuk mengindikasikan level kategori yang berbeda. Untuk pendekatan terhadap analisis skor yang tidak berkategori benar/salah, maka pendekatan politomus merupakan model yang paling memungkinkan

Jika dibandingkan dengan teori klasik, IRT memiliki beberapa kelebihan, yaitu:

1. Item statistics independent terhadap sampel yang diestimasi
2. Skor examinee independent terhadap tingkat kesulitan tes
3. Analisis item mengakomodasi matching antara item test dengan level pengetahuan examinee
4. Analisis tes tidak memerlukan test parallel untuk mengetahui reliability
5. Item statistics dan ability examinee, keduanya dilaporkan dalam skala yang sama

Di bawah asumsi penggunaan model IRT lebih baik dibanding teori tes klasik. Model IRT memerlukan sampel yang besar untuk memperoleh estimasi parameter stabil dan akurat, meski model pengukuran Rasch dapat dipergunakan untuk ukuran sampel kecil. Sebagai konsekuensi, pilihan dari suatu model dapat tergantung pada contoh yang tersedia, terutama sekali di dalam tahap praktek pengujian di lapangan.

PEMBAHASAN

Pembelajaran busana secara umum mempunyai pengertian menciptakan atau membuat suatu busana, baik busana wanita, busana pria, busana anak dengan memperhatikan model, bahan, pola yang digunakan, hiasan, dan teknik menjahitnya. Untuk bisa menciptakan suatu produk busana, memerlukan tahap-tahap yang harus dilakukan. Tahap-tahap yang harus dilakukan apabila akan membuat busana, yaitu: 1) membuat disain busana, 2) membuat pola sesuai dengan ukuran dan model, 3) menjahit busana, yang berkenaan dengan teknologi menjahit, 4) menghias busana (yaitu membuat hiasan pada busana tersebut) (Pori Muliawan,1988)

Tujuan program keahlian Tata Busana membekali peserta didik dengan keterampilan, pengetahuan, dan sikap kompeten dalam:

1. Mengukur, membuat pola, menjahit dan menyelesaikan busana
2. Memilih bahan tekstil dan bahan pembantu secara tepat.
3. Menggambar macam-macam busana sesuai kesempatan.
4. Menghias busana sesuai desain
5. Mengelola usaha di bidang busana.

Model IRT Politomus

Banyak instrumen pengukuran, khususnya instrument untuk mengukur kinerja atau prestasi siswa pada mata pelajaran tertentu, memerlukan respon lebih dari dua kategori. Suatu alasan pengembangan tes yang memerlukan respon lebih dari dua kategori adalah tes tersebut memberikan informasi lebih banyak dibandingkan dengan tes

yang diskor secara dikotomus. Berdasarkan data respon item yang diskor > 2 kategori tersebut, model teori respon item politomus diperlukan untuk menggambarkan hubungan yang tidak linier antara peserta tes dengan kemampuan θ dan probabilitas peserta tes menjawab respon itu pada kategori tertentu.tes yang diukur menggunakan skala unidimensi.

Banyak model yang menggunakan pensekoran politomus, diantaranya adalah: the Graded Response Model (GRM), M-GRM, Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), RSM, dan NRMm (Embretson & Reise, 2000). Semua model tersebut mendasarkan pada asumsi bahwa respon (jawaban) peserta tes pada suatu item tes tergantung kemampuan peserta

Bagaimanapun dengan menggunakan program komputer, pada teori respon butir (IRT) telah berkembang menggunakan pengukuran skala besar. Riset terbaru memfokuskan penggunaan PCM dan GRM polytomous item response theory (PIRT) untuk tes prestasi dengan skor item politomus. Seperti halnya model dikotomus, model politomus digunakan untuk memperoleh kemampuan examinee dan estimasi parameter

Partial Credit Model (PCM) dari Muraki (1992) dan Graded Response Model (GRM) dari Samejima (1969), adalah versi penggunaan tes yang disamakan untuk dua parameter model IRT yang diskor secara polytomous. Bagaimanapun, kedua model tersebut digunakan untuk mengembangkan kemungkinan fungsi pada tes yang diskor secara polytomous. Sementara beberapa riset sudah menerapkan parameter model polytomous (Fitzpatrick, Link, Yet, Burket, Ito, &Sykes, 1996; Maydeu-Olivares, Drasgow, &Mead, 1994; Reise &Yu, 1990). Peneliti mengakui perlunya studi lebih lanjut di dalam bidang polytomus. Reise dan Yu (1990) mengadakan riset tentang penemuan parameter di dalam GRM. Menggunakan simulasi data akan diperoleh estimasi kemampuan, kecenderungan, dan threshold. Bagaimanapun, studi ini dibatasi oleh pemakaian panjangnya test penggunaan kategori respon.

Berdasarkan hasil penelitian Reise dan Yu (1990), ada tiga tujuan penilaian menggunakan test *polytomously-scored*.

1. Pengaruh nomor kategori respon dan panjang tes.
2. Membandingkan ketelitian dari kedua model untuk PCM dan GRM
3. Pengaruh dari kesalahan rater pada PCM dan GRM. Hasil-hasil penelitian akan memberikan petunjuk bagi praktisi dan peneliti mengenai ketelitian dari tiap model seperti respon yang berbeda pada kategori, panjangnya test, dan ketelitian rater di dalam proses penilaian. Selanjutnya dibahas ikhtisar dari PCM dan GRM (Boughton, Klinge, Gierl, 2001)

Sejak 20 tahun yang lalu, *performance test* diperluas menggunakan model politomus. Dengan menggunakan program komputer, pada teori respon butir (IRT) telah berkembang pengukuran skala besar. Riset terbaru memfokuskan penggunaan *Partial Credit Model*

(PCM) dan *Graded Response Model* (GRM) politomous item response theory (PIRT) untuk tes prestasi menggunakan skor politomus. Seperti halnya model dikotomus, model politomus digunakan untuk memperoleh kemampuan *examinee* dan estimasi item parameter. Donoghue, dkk (2000) menyatakan bahwa beberapa teknik dikembangkan untuk menilai kinerja menggunakan program computer. Namun hingga sekarang riset yang menerapkan dan membandingkan dua model tersebut masih sedikit (Keith dkk, 2001). Namun demikian, secara luas respon butir digunakan dengan ekstensif di Canada untuk Penilaian Berkomunikasi di Inggris Columbia (1993) and the School Achievement Indicators Project (Council of Ministers of Education, 1996) and in the United States for the Goals Assessment program (Psychological Corporation,1994), the Stanford Achievement Test, the Golden State Exams, the Illinois Goal Assessment Program (IGAP), the Portfolio Assessment in Vermont, and the Core Assessment CRT Program (Council of Chief State School Officers, U.S.A., 1998, Council of Ministers of Education, Canada,1999).

Partial Credit Model (PCM)

Walaupun terdapat berbagai model polytomous IRT yang berbeda, ,Master (1982) mengemukakan bahwa PCM merupakan salah satu model yang lebih umum digunakan pada model polytomous (Verhelst & Verstralen, 1997). Partial Credit Model (PCM) (Masters, 1982) merupakan bentuk khusus dari IRT satu parameter (Rasch model) dalam kasus dimana skor polytomous Partial Credit Model atau disingkat (PCM) (Masters, 1982) diskor secara polytomous untuk memberi penghargaan terhadap respon, sesuai dengan makna credit partial. Fungsi probabilitas skor didalam kategori x di item i, ability examinee θ, untuk PCM dapat didefinisikan sebagai berikut:

$$P_{ix}\theta = \frac{\exp[\sum_{k=0}^{x}(\theta - b_{ik})]}{\sum_{k=0}^{m_i}\exp[\sum_{k=0}^{k}(\theta - b_{ik})]}$$

dimana :

m_i adalah jumlah katrgori skor minus satu

b_{ik} adalah parameter tingkat kesulitan yang diasosiasikan dengan skor kategori x

Jika dikaitkan dengan pengukuran pada bidang pendidikan, Partial Credit Model (PCM) dapat didisain untuk digunakan pada analisis kepribadian, sikap, prestasi pada pendidikan dan item lain yang menimbulkan respon yang diskor dalam kategori berjenjang, yakni untuk respon item yang bisa diberikan penghargaan secara parsial tidak hanya skor betul dan salah.

Ferara dan Walker (1989) mengambil pendapat para pakar yang telah melakukan pengujian menggunakan PCM, menyatakan bahwa kalibrasi item dengan PCM dapat dikaitkan dengan beberapa nilai tingkat kesulitan, masing-masing mengindikasikan tingkat kesulitan yang diwakili oleh suatu respon atau jawaban examine. Seperti model Rasch untuk item dikotomous, PCM dapat digunakan untuk mengestimasi

ability, kurva item dan karakteristik tes, fungsi item dan informasi tes, fit statistik, standard errors yang berbeda sesuai level ability, dan konstanta linking dapat dihitung dan digunakan untuk item dan tes evaluasi.

Kedua pakar tersebut juga menyatakan bahwa, aplikasi-aplikasi sebelumnya dari PCM pada tes prestasi dalam dunia pendidikan, assessment, merekomendasikan bahwa model PCM dapat dipraktekkan untuk bank soal yang berkaitan dengan problem solving matematika (Masters, 1984), pengukuran sikap dengan computer adaptive (Koch and Doot, 1985), pengembangan item yang berkaitan dengan teori pemerolehan kata (Smith, 1986), equating format-format pada tes essay (Phillips, 1987), equating dengan assessment langsung dan tidak langsung (Phillips, Meadm & Ryan, 1983), mengkonstruksi tugas-tugas naratif (Harris, Laan, & Mossenson, 1988), dan kalibrasi beberapa tipe tugas penulisan (Pollitt & Hutchinson, 1987).

CAT (Computer Adaptive Testing) dapat digeneralisasikan dengan butir yang menggunakan sistem skor partial credit model, yang memungkinkan dibangunnya tes berdasarkan area outcome yang lebih kompleks dari pada hanya jawaban benar dan salah. Respon data untuk kalibrasi menggunakan program komputer berdasarkan pada jumlah sampel besar. Sebelum dikalibrasi perlu diperhatikan persyaratan minimal sampel, untuk memperoleh estimasi parameter yang akurat. Program yang digunakan untuk kalibrasi adalah PARSCALE (Version 3.0) (Gorin, Dood, Fitzpatrick and Yann, 2005). PARSCALE version 3.0 digunakan untuk mengestimasi item parameter berdasar respon data sampel.

Graded Response Model (GRM)

Model GRM dari Samejima (1997), cocok diterapkan apabila urutan respon atau jawaban dari peserta tes saling bergantung, artinya respon kedua dari sebuah item bergantung pada kebenaran dari respon sebelumnya (pertama). Disamping itu pada GRM tidak terdapat parameter guesing (kira-kira). GRM tepat digunakan ketika respon peserta tes terhadap item dapat digolongkan sebagai respon kategori yang berurutan dan tingkat penyelesaian cenderung meningkat, seperti yang ada pada skala likert. GRM adalah generalisasi dari model logistic 2 parameter (2-PL) pada model teori respon item dikotomus.

Secara khusus, jumlah sampel akan mempengaruhi karakteristik empiris pada aplikasi model GRM. Penerapan GRM yang ideal dengan memperhatikan karakteristik model, yaitu ukuran sampel, jumlah item, dan distribusi form. Jawaban mereka akan menimbulkan batasan pada ketepatan estimasi.

Model GRM dari Samejima (1969, 1972, 1997) menggunakan 2 parameter dikotomous untuk menentukan fungsi probabilitas. Hal Ini diciptaka secara menyeluruh fungsi kementakan untuk masing-masing batas kategori. Sebagai contoh, dengan satu item mempunyai kemungkinan empat score , fungsi

dichotomous yang pertama adalah 0 vs 1, 2, atau 3; fungsi kedua adalah 0 atau 1 vs 2 atau3; fungsi yang ketiga adalah 0, 1, atau 2 vs 3. Untuk item dengan kategori k, k-1 fungsi probabilitas yang akan dihitung dengan rumus:

$$P_{jk}\ (\theta) = \frac{exp[Daj(\theta - bjk)]}{1+exp[Daj(\theta - bjk)]}$$

di mana D= 1,7, aj adalah parameter keserongan untuk item j, bjk adalah kategori item parameter threshold untuk kategori k dan item j.

Peran Rater dalam penilaian politomus

Rater atau penilai memegang peran penting dalam penilaian politomus, misalnya pada pengukuran unjuk kerja (*performance assessment*), secara relatif menghasilkan konsistensi yang rendah. *Random error* dari penilai akan mempengaruhi perbedaan skor peserta tes secara keseluruhan. Terdapat tiga sumber kesalahan dalam pensekoran penilaian keterampilan, yaitu: 1) permasalahan instrument, 2) permasalahan prosedural, dan 3) permasalahan pensekoran yang bias. Untuk memperoleh hasil atau skor peserta didik yang sebenarnya, maka konsistensi *inter-rater* sangat diperlukan dalam mendisain rubrik pensekoran secara tepat, pemilihan dan pelatihan penskor (*rater)*, dan meninjau kembali penilai (*rechecking rater performance*). Berdasar penelitian Wainer dan Thissen (1993) bahwa salah satu isu yang berkelanjutan di dalam penilaian-penilaian yang berbasis kinerja adalah tidak adanya kendalan skala karena rater.

Hasil penelitian yang dilakukan oleh Wilson & Case (1997) bahwa penilaian IRT berbeda dengan penilaian secara tradisional. Penilaian secara IRT atau secara politomus, memerlukan ketelitian dari rater atau penilai. Hasil penilaian yang akurat ditentukan kemampuan dan watak penilai atau *rater.* Selanjutnya berdasar pernyataan dari Depdiknas (2007), *rater* keterampilan dilakukan dengan mengamati langsung cara siswa atau yang diuji melakukan pekerjaannya, sehingga *rater* dan yang dinilai berhadapan langsung. Hasil penilaian keterampilan sering dipengaruhi oleh karakteristik rater. Untuk menghindari kesalahan pengukuran yang besar, penilaian dilakukan oleh lebih dari satu orang sebagai suatu tim, masing-masing menilai hal yang sama. Hasil penilaian dari masing-masing *rater* dibandingkan untuk mengetahui konsistensinya. Permasalahan yang sering muncul dalam mendisain dan menggunakan *performance assessment* adalah permasalahan *validity, reliability*, dan *fairness* (Depdiknas,2007). Salah satu isu-isu yang berkelanjutan di dalam penilaian berbasis kinerja adalah tidak adanya keandalan skala karena raters (Thissen, 1993).

Berdasarkan permasalahan yang sering muncul pada penilaian keterampilan bersumber pada rater, maka diperlukan jalan keluar untuk mengatasinya. Hal ini perlu ditempuh untuk mendapatkan kemampuan dan keterampilan peserta tes yang akurat dan sebenarnya, sesuai yang

diharapkan. Penerapan model Partial redit Model (PCM) dan Graded Response Model (GRM), diharapkan memperoleh ketepatan proses menilai untuk memperkecil kesalahan *rater* dan untuk memperoleh estimasi parameter yang akurat. Termasuk juga dalam membuat rubrik perlu tergambar jelas. Disamping itu pelatihan perlu ditingkatkan untuk *rater*, dan pemantauan berkesinambungan dalam proses menilai.

Perbandingan Estimasi Parameter Model PCM vs GRM

Untuk mengetahui ketelitian estimasi parameter, nilai root mean square error (RMSE) antara skor sesungguhnya dan yang diestimasi dihitung dan dibandingkan pada kondisi dan model yang berbeda. RMSE yang kecil menandai (adanya) kesesuaian antara estimasi dan true parameter.

Data dari tes yang diskor secara politomus, dibangkitkan dengan program komputer RESGEN 3 (Muraki, 1999) untuk menetapkan ciri unidimensional latent normal dan polytomous item respon untuk PCM dan GRM. Berdasarkan penemuan Reise dan Yu (1990), digunakan sampel 2000 untuk mendapatkan kestabilan estimasi parameter. Model PCM vs GRM, digunakan kategori respon (4, 6, dan 8 kategori), nomor butir (4, 8, dan 16 butir).

Berdasarkan hasll penelitian yang dilakukan Bougthon (2001), diasin yang dibuat untuk mendapatkan kestabilan estimasi parameter, yakni: kesalahan ketidakhadiran penilai dan kesalahan kehadiran penilai.Untuk memperoleh kestabilan estimasi parameter, dengan pengulangan 30 kali. Kondisi ini dipilih untuk mewakili, menunjukkan jenis-jenis dari variabel-variabel (yaitu., 4, 6, dan 8 kategori untuk 4, 8, dan 16 butir), (cf. *Council of Chief State School Officers*,1998).

Program komputer PARSCALE (Muraki & Bock, 1997) digunakan untuk menaksir parameter kemampuan dan item menggunakan vektor respon examine dihasilkan dengan RESGEN. Pengaturan-pengaturan asumsi digunakan di PARSCALE. Hasil analisis akan diketahui Indek korelasi antara tanpa kesalahan dan dengan kesalahan . PARSCALE kemudian digunakan untuk menaksir parameter kemampuan dan item dari respon yang dimodifikasi vektor-vektor. Dari PARSCALE untuk masing-masing model Parameter-menggunakan suatu metoda transformasi linear:

$$l_x(y) = x = s(X)\left[\frac{y - x(Y)}{s(Y)}\right] + x(X)$$

di mana $l_x(Y)$ adalah estimasi true parameter; x(Y) , x(X)dan s(Y) , s(X) adalah rata-rata dan estimasi simpangan baku dan true parameter (Kolen &Brennan, 1996).

Untuk mengetahui ketelitian estimasi parameter, nilai root mean square error (RMSE) antara skor sesungguhnya dan yang diestimasi dihitung dan dibandingkan pada kondisi dan model yang berbeda, yaitu antara PCM dan GRM. RMSE yang kecil menandai (adanya) kesesuaian antara estimasi dan true parameter.

Berdasarkan hasil, ada faktor dominan pekerjaan kepuasan yang diukur oleh skala dibenarkan penggunaan model yang GRM. GRM yang layak untuk ini adalah data dan perkiraan item nilai parameter yang digunakan dalam simulasi yang diikuti. Data kemudian simulasi dari sekarang ini "populasi parameter" hanya didasarkan pada nilai respon model untuk semua simulasi kondisi kami. Oleh karena itu, dalam simulasi data yang dimaksudkan untuk mewakili satu skala yang mengukur unidimensional (dalam arti IRT) akan secara teoritis sesuai dengan nilai respon model. Simulasi data Kami diperoleh penduduk GRM item parameter untuk simulasi dengan menggunakan MULTILOG 7 (Thissen, 1991)

Untuk mengetahui besarnya nilai Root Mean Square Error (RMSE), baik pada PCM maupun GRM, dilihat dari panjang tes. Berdasarkan hasil penemuan Bougthon (2001), bahwa apabila rater dalam menilai tanpa kondisi error, maka secara umum nilai RMSE akan rendah untuk kedua model, walaupun GRM sedikit lebih tinggi. Nilai RMSE bisa ditekan apabila jumlah butir ditingkatkan atau ditambah, sekaligus menunjukkan estimasi a-parameter kedua model. Apabila nilai RMSE tinggi menunjukkan bahwa beberapa siswa bahkan sebagian besar siswa memperoleh score-score yang ekstrim. Sebaliknya nilai RMSE akan terlihat tinggi, apabila rater atau penilai dalam kondisi error. Dengan demikian ditinjau dari kondisi error dan tanpa kondisi error, akan berpengaruh terhadap nilai RMSE, disamping panjang tes.

Nilai RMSE dalam kondisi error lebih tinggi dibandingkan dengan tanpa kondisi error. Nilai RMSE rendah untuk estimasi parameter kemampuan pada PCM dibandingkan dengan GRM. Pada kategori yang ekstrim atau rendah, maka nilai RMSE akan cenderung tinggi. Penelitian dari Hulin, Lissak, &Drasgow, 1982, menyarankan kasus dari data polytomous, banyaknya penempuh ujian yang diperlukan untuk parameter yang akurat penilaian ,didasarkan pada banyaknya materi, nomor dari kategori-kategori score, dan banyaknya penempuh ujian pada setiap kategori score .

SIMPULAN

Dalam dua dekade terakhir, IRT sebagian besar memfokuskan pada model dichotomous, meskipun estimasi butir diperluas dan penilaian berbasis kinerja menerapkan PCM dan GRM butir polytomously-scored. Sampel yang digunakan di dalam ujian jumlahnya besar. dengan meningkatkan butir polytomously-scored di dalam menguji, perlu menguji PIRT model di bawah kondisi-kondisi ujian yang berbeda. Di dalam ketidakhadiran dari kesalahan rater, GRM dan GPCM estimasi parameter item dan kemampuan ditunjukkan dengan estimasi item sedikit atasan.

Meningkatkan banyaknya butir adalah cara terbaik untuk memperbaiki kemampuan estimasi, selama estimasi parameter relatif stabil dengan butir atau kategori-kategori yang ditingkatkan. Dengan demikian, untuk memperkecil kesalahan rater dan

untuk memperoleh estimasi parameter yang akurat, perlu diperhatikandengan saksama ketepatan proses menilai. Termasuk rubrik-rubrik yang disusun tergambar jelas, pelatihan yang ditingkatkan untuk raters, dan pemantauan berkesinambungan proses membuat angka. Ketelitian dari PCM dan GRM, diperoleh melalui efektivitas raters dan kemampuan rater. Sehingga perlu pemantauan berkesinambungan dalam proses menilai, pembuatan rubrik tergambar dengan jelas, dan pelatihan perlu ditingkatkan untuk rater

REFERENSI

Barbara,J.S.G, Dood, B.G, Fitzpatrick, S.J and Yann. 2005. *Computerized Adaptive Testing With the Partial Credit Model: Estimation Procedures, Population Distributions and Item Pool Characteristics.* Applied Psychological Meassurement: Sage Publications

Bastari. 1998. *An Investigation of Linear and Non Linear Estimates for Multidimensional Graded Response Model.* University of Massachusetts

Boughton, K.A, Klinger, D.A, Gierl, M.J. 2001. *Effects of Random Rater Error on Parameter Recovery of the Generalized Partial Credit Model and Graded Response Model.* Applied Meassurement and Evaluation: University of Alberta. Paper presented: NCME, Seattle, WA. April 10 – 14

Davis, L.L, Dood, B.G. 2005. *Strategies for Controlling Item Exposure in Computerized Adaptive Testing with the Partial Credit Model*

Pearson Educational Measurement. University of Texas at Austin

Demars, C.E, 2003. *Scoring Subscales using Multidimensional Item Response Theory Models.* James Madison University

Ferrara, S, Bartnick, L.W. 1989. *Constructing an Essay Prompt Bank Using the Partial Credit Model.* Maryland State Department of Education

Hol, A.M, Vorst, H.C.M, Mellenbergh, G.J. 2007. *Computerized Adaptive Testing for Polytomous and a Comparison Forms.* Applied Psychological Meassurement. http://apm.sagepub.com

Lautenschlager, G.J, Meade, A.W, Kim, S.H, 2006. *Cautions Regarding Sample Characteristicts When Using the Graded Response Model.* Paper presented at the 21st. Annual Conference of the Society for Industrial and Organizayional Psychology, Dallas, TX

Suchmacker. 2005. Item Response Theory. Applied Meassurement Associates

Wilson, M, Case, H. 1997. *An Examination of Variation in Rater Severity Over Time: A Study in Rater Drift.* Berkeley Evaluation and Assessment Research (BEAR) Center. Berkeley: University of California

Wolfe, E.W. 2004. *Identifying rater affects using latent trait models. Meassurement & Quantitative Methods.* Erickson Hall: Michigan State University