Jurnal Evaluasi Pendidikan, Vol. 2, No. 1, Maret 2011, 18-29

SENSITIVITY OF MANTEL HAENSZEL MODEL AND RASCH MODEL AS VIEWED FROM SAMPLE SIZE

Idrus Alwi
Kanwil Kementerian Agama DKI Jakarta
Jl. D.I Panjaitan No. 10, Jakarta
benmashoor@yahoo.com

Abstract

The aims of this research is to study the sensitivity comparison of Mantel Haenszel and Rasch Model for detection differential item functioning, observed from the sample size. These two differential item functioning (DIF) methods were compared using simulate binary item respon data sets of varying sample size, 200 and 400 examinees were used in the analyses, a detection method of differential item functioning (DIF) based on gender difference. These test conditions were replication 4 times. For both differential item functioning (DIF) detection methods, a test length of 42 items was sufficient for satisfactory differential item functioning (DIF) detection with detection rate increasing as sample size increased. Finding the study revealed that the empirical result show Rasch Model are more sensitive to detection differential item functioning (DIF) than Mantel Haenszel. With reference to findings of this study, it is recomended that the use of Rasch Model in evaluation activities with multiple choice test. For this purpose, it is necessary for every school to have some teachers who are skillfull in analyzing results of test using modern methods (Item Response Theory).

Keywords: *Mantel Haenszel Model, Rasch model, differential item functioning, sample size*

Jurnal Evaluasi Pendidikan, Vol. 2, No. 1, Maret 2011, 19-29

PERBANDINGAN KEPEKAAN MODEL MANTEL HAENSZEL DAN MODEL RASCH DALAM MENDETEKSI KEBERBEDAAN FUNGSI BUTIR DITINJAU DARI UKURAN SAMPEL

Idrus Alwi
Kanwil Kementerian Agama DKI Jakarta
Jl. D.I Panjaitan No. 10 Jakarta
benmashoor@yahoo.com

Abstrak

Penelitian ini bertujuan mengetahui perbandingan kepekaan model Mantel Haenszel dan model Rasch dalam mendeteksi keberbedaan fungsi butir ditinjau dari ukuran sampel. Kedua metode deteksi keberbedaan fungsi butir menggunakan butir soal yang bersifat dikotomi dengan ukuran sampel yang berbeda, yaitu 200 responden dan 400 responden digunakan dalam analisis, sedangkan untuk mendeteksi ada tidaknya keberbedaan fungsi butir berdasarkan pada perbedaan jenis kelamin. Tes ini dilakukan pengulangan sebanyak 4 (empat) kali. Untuk masing-masing metode deteksi keberbedaan fungsi butir digunakan sebanyak 42 butir, dimana penggunaan ukuran sampel yang semakin meningkat mempengaruhi kepekaan metode deteksi keberbedaan fungsi butir yang digunakan. Hasil penelitian menunjukkan bahwa metode deteksi keberbedaan fungsi butir model Rasch memiliki kepekaan yang lebih tinggi dibandingkan dengan metode Mantel Haenszel. Berdasarkan hasil penelitian tersebut, metode deteksi keberbedaan fungsi butir model Rasch dapat digunakan untuk kegiatan-kegiatan evaluasi yang menggunakan butir soal dalam bentuk pilihan ganda. Untuk tujuan ini, sangat penting untuk setiap sekolah memiliki guru-guru yang menguasai dengan baik penggunaan Teori Responsi Butir dalam menganalisis butir soal.

Kata kunci: model Mantel Haenszel, model Rasch, keberbedaan fungsi butir, ukuran responden

PENDAHULUAN

Dalam lingkungan pendidikan di sekolah, informasi tentang keberhasilan guru menyajikan bahan pelajaran serta sejauhmana peserta didik telah menyerap materi yang diajarkan dapat diketahui dari penilaian dan pengukuran. Karena itu, penilaian dan pengukuran merupakan bagian yang penting di dalam proses belajar mengajar. Salah satu alat ukur yang digunakan selama ini untuk memperoleh informasi tersebut adalah tes prestasi belajar berupa tes tertulis.

Tes prestasi merupakan upaya pengukuran terencana yang digunakan oleh guru untuk memberi kesempatan bagi siswa dalam memperlihatkan prestasi mereka yang berkaitan dengan tujuan dan fungsi yang telah ditentukan. Ebel (1979: 21) mengatakan bahwa fungsi utama tes prestasi di kelas adalah mengukur prestasi belajar para siswa. Suatu kesalahpahaman bila menganggap bahwa apa yang dapat dilakukan oleh tes prestasi belajar semata-mata

memberikan angka untuk dimasukkan ke dalam rapor siswa atau ke dalam laporan hasil studi mahasiswa. Sesungguhnya prosedur tes, berguna mengukur prestasi mengandung nilai-nilai pendidikan yang sangat penting.

Untuk mengukur kemampuan teoritis siswa di bidang matematika diperlukan perangkat tes yang berkualitas. Perangkat tes ini harus benar-benar dapat mengukur apa yang seharusnya diukur dan memberikan hasil yang dapat dipercaya. Untuk itu, diperlukan alat ukur yang memiliki tingkat kesulitan yang kurang lebih sepadan dengan kemampuan peserta tes, indeks daya beda yang tinggi, serta faktor tebakan (*guessing*) yang seminimal mungkin sehingga dapat memberikan informasi pengukuran yang akurat. Selain itu, butir tes yang baik harus terbebas dari bias. Tes yang baik tidak memihak pada kelompok tertentu atau golongan tertentu dari peserta tes. Tes yang baik akan memberikan hasil pengukuran yang sama terhadap peserta tes yang memiliki kemampuan sama meskipun berasal dari kelompok atau golongan yang berbeda. Bila tes memberikan hasil yang berbeda maka tes tersebut bias, yang berarti perangkat tersebut tidak valid secara konstruktif. Menurut Surapranata (2004: 46) soal yang bias adalah soal yang membedakan kelompok. Zumbo (1999: 12) mengatakan bahwa bias soal terjadi ketika peserta tes dari dua kelompok dapat menjawab benar yang lebih sedikit dari kelompok lainnya. Holland dan Wainer (1993: 4) mengatakan pada dasarnya sama dalam memberikan definisi tentang bias soal adalah bila individu yang mempunyai kesamaan dalam kemampuan yang termasuk dalam kelompok yang berbeda, tidak mendapatkan kesamaan kemungkinan dalam menjawab soal dengan benar. Menurut Hambleton dan Swaminathan (1994: 282) pengertian dari butir yang bias adalah adanya perbedaan skor perolehan yang disebabkan oleh adanya unsur yang menguntungkan atau merugikan bagi peserta sehingga tingkat kesukaran bagi kelompok peserta tes berbeda.

Adapun faktor-faktor yang menyebabkan terjadinya bias butir soal dalam pelaksanaan tes adalah pengaruh perbedaan ras, jenis kelamin, wilayah, budaya dan etnis. Berk (1982: 1) mengatakan bahwa penyebab adanya bias butir soal adalah jenis kelamin, ras dan etnis.

Perbedaan skor tes tersebut disebabkan oleh faktor-faktor perbedaan ras, jenis kelamin, atau perbedaan etnis. Untuk mengetahui ada tidaknya bias pada suatu butir diperlukan analisis keberbedaan fungsi butir atau *Differential Item Functioning (DIF*). Menurut Perrone (2006: 3) *differential item functioning* adalah kumpulan metode statistik yang digunakan untuk menentukan apakah butir-butir soal sudah sesuai dan adil untuk menguji pengetahuan dari berbagai kelompok ujian (misalnya, laki-laki vs perempuan atau Kaukasus vs Afrika-Amerika). *Differential item functioning* akan terjadi bila terdapat perbedaan yang signifikan secara statistik terhadap kemungkinan hasil tes dari dua kelompok (misalnya, laki-laki dan perempuan), yang memiliki kemampuan yang sama tetapi

menunjukkan perbedaan kemungkinan dalam menjawab butir soal. Wieberg (2007: 1) menyatakan d*ifferential item functioning* (*DIF*) adalah perbedaan yang tidak diharapkan diantara beberapa kelompok ujian yang seharusnya hasil ujian tersebut sebanding berdasarkan atribut yang diukur oleh butir soal dan tes yang dikerjakan.

Menurut Master dan Keeves (1999: 221-232) terdapat dua pendekatan analisis data yang digunakan dalam menentukan ada atau tidaknya keberbedaan fungsi butir dari sebuah butir soal, yaitu dengan pendekatan klasik dan pendekatan teori responsi butir. Deteksi kebiasan soal dengan pendekatan klasik dilakukan dengan metode transformasi tingkat kesukaran soal, daya beda soal, pendekatan tabel kontigensi (model Mantel Haenszel), standarisasi prosedur, metode registrasi logistik, dan analisis distraktor. Sedangkan pendekatan teori responsi butir dilakukan dengan model Rasch (satu parameter), dua parameter dan tiga parameter.

Mantel dan Haenszel menampilkan model untuk suatu studi pemadanan kelompok, yang oleh Holland dan Thayer dipakai untuk mendeteksi *DIF,* yang kemudian terkenal dengan metode Mantel Haenszel. Menurut Roussos, Schnipke dan Pashley (2000: 5) metode ini merupakan metode yang *powerful* dan digunakan di *Educational Testing Service (ETS)* di Amerika Serikat. Seluruh metode dan teknik yang telah dikembangkan untuk mengidentifikasi keberbedaan fungsi butir mengambil asumsi yang sama, yaitu kelompok butir atau pengujian yang berisi butir bersifat homogen dan satu dimensi. Asumsi ini berlaku untuk model IRT dan Mantel Haenszel.

Model Rasch maupun model Mantel Haesnzel memerlukan ukuran sampel yang sesuai sehingga diperoleh tingkat kepekaan yang tinggi. Hasil penelitian yang dilakukan oleh Bezruczko (1989: 32) tentang kestabilan metode deteksi bias soal antara model Rasch dan model Mantel Haenszel dengan menggunakan ukuran sampel 300 dan 1000, tingkat kestabilan model Rasch dan model Mantel Haenszel lebih tinggi dengan menggunakan ukuran sampel 1000 dibandingkan dengan ukuran sampel 300. Secara umum, ukuran sampel menjadi salah satu hal penting dalam pelaksanaan pendeteksian *DIF*.

Berdasarkan uraian di atas, timbul pertanyaan manakah yang lebih peka model Mantel-Haenszel atau model Rasch dalam mendeteksi keberbedaan fungsi butir ditinjau ukuran sampel 200 dan 400 atau dua kali dari responden semula. Tujuan dari penelitian adalah untuk mengetahui kepekaan metode deteksi keberbedaan fungsi butir model Rasch dibandingkan dengan metode deteksi keberbedaan fungsi butir prosedur Mantel Haenszel ditinjau dari ukuran sampel.

METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini adalah eksperimen dengan rancangan *treatment by level design*. Variabel penelitian dibedakan menjadi variabel bebas dan variabel terikat. Variabel terikat dari

penelitian ini adalah kepekaan deteksi keberbedaan fungsi butir, dan variabel bebas penelitian ini adalah metode deteksi keberbedaan fungsi butir yang dibedakan menjadi dua kategori, yaitu prosedur Mantel Haenszel dan model Rasch dan ukuran sampel yang dibedakan menjadi 200 dan 400 orang. Untuk kemudahan menghitung nilai rata-rata dan deviasi standar masing-masing sel, dilakukan replikasi sebanyak 4 kali. Populasi penelitian ini adalah butir tes bentuk pilihan ganda dengan sub-stansi materi bahan ajar mata pelajaran matematika Madrasah Tsanawiyah. Sampel adalah 50 butir tes mata pelajaran matematika bentuk pilihan ganda. Responden sebagai peserta tes yaitu siswa Madrasah Tsanawiyah Negeri kelas III di Jakarta Timur, yang secara keseluruhan siswa peserta tes berjumlah 1000 orang.

Metode deteksi keberbedaan fungsi butir yang digunakan dalam penelitian ini terbagi menjadi dua, yaitu metode deteksi keberbedaan fungsi butir dengan model Mantel Haenszel dan metode deteksi keberbedaan fungsi butir dengan model Rasch. Model Mantel Haenszel adalah rumus yang dipakai untuk menganalisis butir-butir tes pilihan ganda untuk mata pelajaran matematika di kelas III. Rumus ini mengklasifikasikan siswa yang menjawab tes menjadi dua kelompok yaitu kelompok fokus dan kelompok referensi, sedangkan model Rasch adalah rumus yang dipakai untuk menganalisis butir-butir tes pilihan ganda untuk mata pelajaran matematika di kelas III. Hasil perhitungan dari rumus ini diperoleh dari perbedaan perhitungan parameter tingkat kesukaran antara kelompok pria dan kelompok wanita yang selanjutnya selisih hasil perhitungan tersebut dibagi dengan akar dari masing-masing *standart error* kedua kelompok.

Kisi-kisi instrumen sebelum diuji coba terdiri dari 50 butir tes. Kisi-kisi instrumen mencakup variabel kemampuan kognitif siswa. Untuk mengetahui daya beda dan tingkat kesukaran butir digunakan analisis butir klasik dengan program *Iteman*. Hasil analisis butir yang diperoleh dijadikan instrumen dalam mengumpulkan data selanjutnya. Instrumen sebelum digunakan untuk pengukuran variabel kemampuan kognitif siswa dilakukan ujicoba terlebih dahulu. Ujicoba instrumen dilaksanakan dengan cara memeriksa validitas butir tes, yaitu kesesuaian antara setiap butir dengan materi pelajaran, indikator, dan dimensi variabel yang diukur. Pemeriksaan dilakukan oleh para guru matematika di Madrasah Tsanawiyah yang sudah cukup lama mengajar yaitu 5 tahun lebih. Jumlah guru matematika yang digunakan sebagai validator adalah 20 orang. Perhitungan reliabilitas instrumen dalam penelitian ini dilakukan dengan Alpha Cronbach.

Teknis analisis data untuk menentukan ada tidaknya keberbedaan fungsi butir dari model Rasch digunakan program Rasch (*Rasch Model Item Calibration Program*) yang dikembangkan oleh Benjamin D. Wright. Sedangkan penentuan ada tidaknya keberbedaan fungsi butir dari prosedur Mantel Haenszel digunakan program Mantel Haenszel-Cochran. Teknik uji statistik yang digunakan untuk menguji hipotesis yaitu analisis varian dua jalan.

HASIL PENELITIAN

Uji persyaratan analisis dan deskripsi data berdasarkan hasil uji validitas yang dilakukan oleh para pakar diperoleh nilai reliabilitas interrater sebesar 0,90. Sedangkan hasil uji empiris pada butir soal tes pilihan ganda diperoleh 42 butir soal tes pilihan ganda yang berasal dari seleksi hasil uji coba 50 butir soal. Berdasarkan kriteria yang telah ditetapkan, maka diperoleh 42 butir soal yang memenuhi kriteria sebagai instrumen penelitian.

Pengujian unidimensional bertujuan untuk melihat apakah tes yang telah disusun mengukur satu karakteristik di kalangan peserta. Berdasarkan hasil hitung kelompok responden 200 memiliki nilai *eigen value* faktor pertama 217,5 kali dari faktor kedua. Kelompok responden 400 memiliki nilai *eigen value* faktor pertama 73,15 kali dari faktor kedua. Ini berarti kedua kelompok tersebut telah memenuhi syarat unidimensi.

Pengujian independensi lokal bertujuan untuk melihat apakah nilai kemungkinan seorang individu menjawab benar suatu soal tidak bergantung pada jawaban soal lainnya dengan kata lain kemampuan peserta dalam satu subpopulasi yang sama independen terhadap butir. Berdasarkan hasil hitung nilai kovariansi antar interval kemampuan peserta adalah kecil dan mendekati nol, ini dapat disimpulkan bahwa persyaratan *local independence* telah terpenuhi.

Pengujian kecocokan butir bertujuan untuk melihat apakah butir-butir yang digunakan sesuai dengan model yang dipakai yaitu model Rasch. Hasil hitung menggambarkan bahwa semua butir cocok dengan model yang digunakan, yaitu model Rasch. Pemeriksaan invariansi kelompok bertujuan untuk melihat apakah semua subkelompok memiliki karakteristik butir yang sama walaupun kelompok peserta yang menjawab butir yang sama itu berubah-ubah, sehingga diketahui bahwa subkelompok homogen apabila semua responden dalam subkelompok itu memiliki kemampuan sama. Hasil hitung memperlihatkan bahwa hasil pengujian menunjukkan nilai korelasi *Product Moment Pearson* dari kedua kelompok cukup tinggi yaitu 0,901 untuk ukuran sampel 200 dan 0,913 untuk ukuran sampel 400. Ini berarti dapat disimpulkan bahwa persyaratan invariansi kelompok terpenuhi.

Hasil deteksi keberbedaan fungsi butir pada ukuran sampel 400 dengan menggunakan model Mantel Haenszel dari 42 butir soal matematika tingkat SMP/MTs yang terdeteksi mengandung keberbedaan fungsi butir adalah 4 butir, yaitu butir 4, 12, 41 dan 42. Sedangkan hasil pendeteksian model Mantel Haenszel pada ukuran sampel 200 terdeteksi mengandung keberbedaan fungsi butir adalah 2 butir, yaitu butir nomor 41 dan 42. Hasil deteksi keberbedaan fungsi butir pada ukuran sampel 400 dengan menggunakan model Rasch dari 42 butir soal matematika tingkat Madrasah Tsanawiyah yang terdeteksi mengandung keberbedaan fungsi butir adalah 14 butir, yaitu butir nomor 2, 4, 11, 12, 15, 16, 28, 30, 33, 36, 37, 40, 41 dan 42. Sedangkan hasil pendeteksian model Rasch pada ukuran sampel 200 terdeteksi mengandung keberbedaan

fungsi butir adalah 12 butir, yaitu butir nomor 2, 4, 12, 16, 28, 30, 33, 36, 37, 40, 41 dan 42.

Butir-butir yang terdentifikasi mengandung keberbedaan fungsi butir dengan menggunakan metode Mantel Haenzel dan metode Rasch sebagai berikut:

Tabel 1. Butir yang Terindikasi Keberbedaan Fungsi Butir

Replikasi	Ukuran Responden			
	Mantel Haenszel		Rasch	
	200 Responden	400 Responden	200 Responden	400 Responden
Penelitian dasar	41, 22	4, 12, 41, 42	2, 4, 12, 16, 28, 30, 33, 36, 37, 40, 41, 42	2, 4, 11, 12, 15, 16, 28, 30, 33, 36, 37, 40, 41, 42
Replikasi ke-1	4, 12	4, 12, 41, 42	2, 14, 12, 15, 16, 36, 37, 40, 41, 42	2, 14, 11, 12, 15, 16, 28, 30, 33, 36, 37, 40, 41, 42
Replikasi ke-2	4, 12, 41, 40	4, 12, 41, 42	2, 14, 12, 15, 36, 40, 41, 42	2, 14, 12, 15, 30, 36, 37, 40, 41, 42
Replikasi ke-3	12, 41, 42	2, 4, 12, 40, 41, 42	2, 4, 12, 15, 16, 36, 37, 40, 41, 42	2, 4, 12, 15, 16, 28, 30, 36, 37, 40, 41, 42
Replikasi ke-4	41, 42	2, 4, 41, 42	2, 4, 12, 15, 36, 40, 41, 42	2, 4, 12, 15, 16, 30, 36, 37, 40, 41, 42

Pengujian hipotesis pertama, yaitu hipotesis tentang perbedaan kepekaan metode deteksi keberbedaan fungsi butir model Rasch lebih tinggi daripada model Mantel-Haenszel. Berdasarkan pada hasil analisis pengujian hipotesis pada tabel 7, harga F_{hitung} antar metode deteksi keberbedaan fungsi butir adalah 121,69 lebih besar dari harga F_{tabel} yaitu 4,49 ini berarti hipotesis nol ditolak. Ka-rena Ho ditolak berarti data mendukung hipotesis. Berdasarkan tabel 6, kepekaan metode deteksi keberbedaan fungsi butir model Rasch lebih tinggi daripada metode deteksi keberbedaan fungsi butir model Mantel Haenszel.

Pengujian hipotesis kedua, yaitu hipotesis tentang perbedaan kepekaan metode deteksi keberbedaan fungsi butir model Rasch dan model Mantel Haenszel dengan ukuran sampel 400 lebih tinggi daripada ukuran sampel 200. Berdasarkan pada hasil analisis pengujian hipotesis pada tabel 7, harga F_{hitung} antar ukuran sampel adalah 7,20 lebih besar dari harga F_{tabel} yaitu 4,49 ini berarti hipotesis nol ditolak. Karena Ho ditolak berarti data mendukung hipotesis. Berdasarkan tabel 1 dan tabel 2, kepekaan metode deteksi keberbedaan fungsi

butir model Rasch dan model Mantel Haenszel dengan ukuran sampel 400 lebih tinggi daripada ukuran sampel 200.

Tabel 2. Hasil Uji Anova Dua Jalan

Sumber Variasi	JK	Db	RJK	F_{hitung}	$F_{tabel\ (1:19)(0,05)}$
Antar metode	273,8	1	273,8	121,69	
Antar respoden	16,2	1	16,2	7,20	4,49
Interaksi antara Metode x respond.	3,2	1	3,2	1,42	
Dalam	36	16	2,25		
Total	329,2	19			

Tabel 3. Rekapitulasi Butir-Butir yang Terindikasi Keberbedaan Fungsi Butir

Replikasi	Mantel Haenszel		Rasch	
	Resp. 200	Resp. 400	Resp. 200	Resp. 400
1	4 butir	2 butir	12 butir	14 butir
2	2 butir	4 butir	10 butir	14 butir
3	4 butir	4 butir	8 butir	10 butir
4	3 butir	6 butir	10 butir	12 butir
5	2 butir	4 butir	8 butir	11 butir

Pengujian hipotesis ketiga, yaitu hipotesis tentang tingkat kepekaan metode deteksi keberbedaan fungsi butir pada ukuran sampel 200, antara yang mendapat perlakuan dengan metode Rasch dan metode Mantel Haenszel. Berdasarkan pada hasil analisis pengujian hipotesis dengan menggunakan metode Tukey pada tabel 10, harga Q_{hitung} = 9,85 lebih besar dari harga Q_{tabel} = 5,22. Ini berarti hipotesis nol ditolak. Karena Ho ditolak berarti data mendukung hipotesis.

Pengujian hipotesis keempat, yaitu hipotesis tentang tingkat kepekaan metode deteksi keberbedaan fungsi butir pada ukuran sampel 400, antara yang mendapat perlakuan dengan model Rasch dan model Mantel Haenszel. Berdasarkan pada hasil analisis pengujian hipotesis dengan menggunakan uji Tukey, harga Q_{hitung} = 12,24 lebih besar dari harga Q_{tabel} = 5,22. Ini berarti hipotesis nol ditolak. Karena Ho ditolak berarti data mendukung hipotesis. Ini berarti pada ukuran sampel 400, perbedaan kepekaan deteksi keberbedaan fungsi butir model Rasch lebih tinggi dari model Mantel Haenszel. hasil ini didukung banyak butir yang terindikasi keberbedaan fungsi butir pada tabel 2 dan tabel 3 yaitu pada ukuran sampel 200 dan 400, perbedaan kepekaan deteksi keberbedaan fungsi butir model Rasch lebih tinggi secara signifikan dari model Mantel Haenszel.

Pengujian hipotesis kelima, yaitu hipotesis tentang tingkat kepekaan metode deteksi keberbedaan fungsi butir pada model Mantel Haenszel, antara yang diberi perlakuan dengan menggunakan ukuran sampel 400 dengan ukuran

sampel 200. Berdasarkan pada hasil analisis pengujian hipotesis dengan mengguna-kan uji Tukey, harga Q_{hitung} = 1,49 lebih kecil dari harga Q_{tabel} = 5,22. Ini berarti hipotesis nol diterima. Karena Ho diterima berarti data tidak mendukung hipotesis. Ini berarti model Mantel Haenszel, antara yang diberi perlakuan dengan ukuran sampel 400 tidak berbeda dibandingkan dengan ukuran sampel 200.

Pengujian hipotesis keenam, yaitu hipotesis tentang tingkat kepekaan metode deteksi keberbedaan fungsi butir pada model Rasch, antara yang diberi perlakuan dengan menggunakan ukuran sampel 400 dengan ukuran sampel 200. Berdasarkan pada hasil analisis pengujian hipotesis dengan menggunakan uji Tukey, harga Q_{hitung} = 3,88 lebih kecil dari harga Q_{tabel} = 5,22. Ini berarti hipotesis nol diterima. Karena Ho diterima berarti data tidak mendukung hipotesis. Ini berarti pada taraf model Rasch, antara yang diberi perlakuan dengan ukuran sampel 400 tidak berbeda dibandingkan dengan ukuran sampel 200.

PEMBAHASAN

Kepekaan metode deteksi keberbedaan fungsi butir dilihat dari semakin banyaknya butir soal yang terdeteksi mengandung keberbedaan fungsi butir. Sehingga dari hasil penelitian ini dapat dinyatakan bahwa (1) metode deteksi keberbedaan fungsi butir model Rasch memiliki kepekaan yang lebih tinggi dibandingkan dengan metode deteksi keberbedaan fungsi butir model Mantel Haenszel, (2) metode deteksi keberbedaan fungsi butir model Rasch memiliki kepekaan yang lebih tinggi dibandingkan dengan metode deteksi keberbedaan fungsi butir model Mantel Haenszel dengan menggunakan responden 400, (3) metode deteksi keberbedaan fungsi butir model Rasch memiliki kepekaan yang lebih tinggi dibandingkan dengan metode deteksi keberbedaan fungsi butir model Mantel Haenszel dengan menggunakan responden 200. Ini sesuai dengan penelitian yang dilakukan oleh Schulz (1990) menguji tingkat kepekaan antara model Rasch dengan prosedur Mantel Haenszel dalam mendeteksi ada tidaknya bias butir. Hasilnya menunjukkan bahwa metode deteksi bias butir model Rasch memilki tingkat yang lebih tinggi dibandingkan dengan metode deteksi bias butir prosedur Mantel Haenszel.

Hasil penelitian ini menyatakan metode deteksi keberbedaan fungsi butir model Rasch memiliki kepekaan yang lebih tinggi dibandingkan dengan metode deteksi keberbedaan fungsi butir model Mantel Haenszel baik dengan menggunakan responden 400 dan responden 200, hal ini didukung dengan dilakukannya 5 replikasi sehingga secara keseluruhan terdapat 20 kasus yang dipelajari. Bila dilihat dari banyaknya butir yang dideteksi mengandung keberbedaan fungsi butir, maka model Rasch memiliki kepekaan yang lebih tinggi dibandingkan dengan model Mantel Haenszel. Hal ini dapat disebabkan oleh: (1) dalam menganalisis ada tidaknya keberbedaan fungsi butir, model Rasch menghitung tingkat kesukaran butir secara terpisah antara laki-laki dan perempuan, setelah itu dihitung kembali selisih dari masing-masing tingkat

kesukaran, sedangkan model Mantel Haenszel menganalisis butir-butir yang benar dan yang salah sekali jalan, (2) sebelum menganalisis keberbedaan fungsi butir dengan model Rasch diperlukan uji persyaratan seperti uji unidimensi dengan menggunakan faktor analisis, uji independensi lokal dengan menggunakan matrik kovariansi dan uji kecocokan data dengan menggunakan rumus *chi-square*. Sebaliknya model Mantel Haenszel tidak melalui ketiga uji sebagaimana yang digunakan oleh metode Rasch.

Selanjutnya dari hasil analisis, butir yang menguntungkan peserta perempuan karena butir-butir soal tersebut berkaitan dengan materi aljabar. Sedangkan butir berikutnya lebih menguntungkan peserta laki-laki karena butir-butir soal tersebut berkaitan dengan materi geometri. Pelajaran geometri dikategorikan sebagai pengetahuan visual ruang, yang dalam hal ini pihak laki-laki lebih menguntungkan dari pihak perempuan. Laki-laki lebih baik daripada perempuan pada tugas yang memerlukan *transformasi visual spatial working memory*. Berbagai penelitian empirik menunjukkan bahwa laki-laki lebih menguasai geometri sehingga butir-butir tes yang mengandung materi geometri lebih mudah dijawab oleh laki-laki dibandingkan oleh perempuan. Sebaliknya butir-butir tes yang mengandung materi aljabar lebih mudah dijawab oleh perempuan dibandingkan oleh laki-laki.

Hasil penelitian ini dapat dilanjutkan dengan melakukan manipulasi panjang tes, banyak butir gandeng, wilayah dan budaya untuk model deteksi lainya dengan pendekatan teori tes klasik maupun teori tes modern. Selain itu perlu dibuat suatu program deteksi keberbedaan fungsi butir dalam bentuk *softwere*. Pada program tersebut perlu dibuat *softwere* yang berisi metode deteksi keberbedaan fungsi butir secara modern dengan satu kali jalan. Selama ini *softwere* yang tersedia untuk mendeteksi keberbedaan fungsi butir satu kali jalan adalah dengan cara klasik sedangkan cara modern harus beberapa kali memasukan dan menganalisis data ke dalam aplikasi program yang berbeda. program ini dapat membantu sekolah untuk menentukan kualitas butir soal yang disusun pada setiap bidang studi. Selain itu *softwere* ini dapat digunakan oleh Kantor Dinas kementrian Pendidikan Nasional atau Kantor Wilayah Kementrian Agama untuk melihat hasil sumatif secara akurat dalam mendapatkan kualitas tes yang berkualitas.

Deteksi keberbedaan fungsi butir tes hasil belajar di sekolah dasar dan menengah dapat menggunakan Metode Mantel Haenzel atau metode Rash dengan ukuran responden yang kecil.

SIMPULAN

Butir soal dikatakan mengandung keberbedaan fungsi butir jika probabilitas menjawab benar butir soal dari dua kelompok berbeda, padahal mereka mempunyai kemampuan sama. Kelompok peserta tes perempuan merupakan kelompok fokal sedangkan kelompok peserta tes laki-laki merupakan

kelompok referensi. Untuk mengetahui butir soal mengandung keberbedaan fungsi butir atau *differential item functioning* maka dapat menggunakan metode deteksi yang terdiri dari model Rasch dan model Mantel Haenszel dengan memperhatikan ukuran sampel. Hasil penelitian metode deteksi keberbedaan fungsi butir dengan menggunakan metode Rasch dan Mantel Haenzel, didapatkan kesimpulan tingkat kepekaan deteksi keberbedaan fungsi butir model Rasch lebih tinggi dibandingkan dengan metode deteksi keberbedaan fungsi butir model Mantel Haenszel pada ukuran sampel yang berbeda.

DAFTAR PUSTAKA

Berk, Ronald A. (ed.). (1982). *Handbook of Methods For Detecting Test Bias.* New York: The John Hopkins University Press.

Bezruczko, Nikolaus, *et al*. (1989). *The Stability of Four Methods for Estimating Item Bias*. Chicago: Department of Research and Evaluation Chicago Public School.

Ebel, Robert L. (1979). *Essentials of Education Measurement.* New Jork: Prentice-Hall, Inc.

Hambleton, Ronald K. dan Hariharan Swaminathan. (1984). *Item Response Theory, Principles and Applications.*Boston: Kluwer Academic Publisher.

Holland, Paul W. dan Howard Wainer (ed.). (1993). *Differential Item Functioning.* New Jersey: Lawrence Erlbaum Associates Publisher.

Master, Geofferey N. dan John P. Keeves (ed.). (1999). *Advance in Measurement in Education Reseach and Assessment*. United Kingdom: Elsevier Science Ltd.

Perrone, Michael. (2006). *Differential Item Functioning and Item Bias: Critical Considerations in Test Fairness*, 2006, http://journals.tc-library.org/ templates/about/editable/pdf/Perrone%20Forum.pdf.

Roussos, Louis A., Deborah L. Schnipke, dan Peter J Pashley. (2000). "A Formulation Of The Mantel-Haenszel Diferential Item Function Parameter With Practical Implication". *LSAC Statistical Report 96-03*.

Schulzt, E. Matthew, *et al*. (1990). "DIF Detection: Rasch Versus Mantel Haenszel" *Rasch Measurement Transaction*, Vol. 4 (2). http://www.rasch.org/ rmt/ rmt42f.htm.

Surapranata, Sumarna. (2004). *Analisis, Validitas, Reliabilitas dan Interpretasi Hasil Tes.* Bandung: Rosda.

Weinberg, Marie. (2007). *Measuring And Detecting Differential Item Functioning In Criterion Referenced Licesing Test*., http://www8.umu.se/ eadmeas/publikationer/pdf.

Zumbo, Bruno D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning*. Ottawa: Directorate of Human Resources Reseach and Evaluation National Defense Headquarter.