

SENSITIVITY OF DIFFERENTIAL ITEM FUNCTIONING (DIF) DETECTION METHOD

Sudaryono

Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Raharja
Jl. Jend. Sudirman No. 40, Modern Cikokol, Tangerang
sudaryono2@yahoo.com

Abstract

The prime objective of this study is to find out the empirical data about a comparison sensitivity as Scheuneman's Chi-square, Mantel-Haenszel methods and Item Response Theory (IRT) Rasch model in Differential Item Functioning (DIF) detection. The study used experimental methods with 1 x 3 design. The independent variables were, Scheuneman's Chi-Square methods, Mantel-Haenszel methods and IRT Rasch model. This specific research aims at revealing: (1) the characteristics of test items based on the of classical test-theory, (2) the standard error measurement of classical test theory, and (3) the items detected as containing Differential Item Functioning (DIF) based on gender. Analysis of this study was based on testee's response to the Mathematics test in the National Examination at Senior High School in Tangerang, 2008/2009 academic year. The data source was computer answer sheets of 5000 consist of 2500 from male student responses and 2500 female student responses, established using the simple random sampling technique. A result of the descriptive analysis using the classical test theory indicated that 28 out of 40 mathematics test items were good, with a reliability index of 0,827. The results indicated that all methods were good for the detection of DIF, but IRT Rasch model was the most sensitive among those methods compared with Mantel-Haenszel, and Sceaneman's chi-square methods. The result of DIF analysis on the mathematics test was find any items containing DIF based on gender differences.

Keywords: *DIF, Rasch model, Mantel-Haenszel methods, Scheuneman's Chi-square methods, classical test theory, National Examination*

SENSITIVITAS METODE PENDETEKSIAN DIFFERENTIAL ITEM FUNCTIONING (DIF)

Sudaryono

Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Raharja
Jl. Jend. Sudirman No. 40, Modern Cikokol, Tangerang
sudaryono2@yahoo.com

Abstrak

Tujuan utama dari penelitian ini adalah untuk mengetahui data empiris tentang sensitivitas perbandingan antara metode *Chi-Square* Scheuneman, metode Mantel-Haenszel metode teori responsi butir model Rasch dalam mendeteksi keberadaan *Differential Item Functioning (DIF)*. Penelitian ini menggunakan metode eksperimental dengan desain 1 x 3. Variabel independen dalam penelitian ini adalah, *Chi-square* Scheuneman, metode Mantel-Haenszel, dan teori responsi butir model Rasch. Secara spesifik tujuan penelitian ini adalah untuk mengungkapkan: (1) karakteristik butir tes berdasarkan teori tes klasik, (2) kesalahan pengukuran standar berdasarkan teori tes klasik, dan (3) deteksi butir soal yang terindikasi mengandung *Differential Item Functioning (DIF)* berdasarkan perbedaan gender. Analisis data penelitian ini didasarkan pada responsi peserta tes Ujian Nasional Matematika pada Sekolah Menengah Atas di Tangerang tahun akademik 2008/2009. Sumber data penelitian ini adalah lembar jawaban komputer dari 5000 siswa yang terdiri dari 2500 siswa pria dan 2500 siswa perempuan yang diambil secara acak dengan menggunakan teknik *simple random sampling*. Hasil analisis deskriptif dengan menggunakan teori tes klasik menunjukkan bahwa ada 28 butir tes dari 40 butir tes matematika, dengan indeks keandalan reliabilitas sebesar 0,827. Hasil penelitian menunjukkan bahwa semua metode yang digunakan untuk mendeteksi *DIF* cukup baik, namun metode teori responsi butir model Rasch adalah model yang paling sensitif dibandingkan dengan metode Mantel-Haenszel, dan metode *Chi-square* Scheuneman.

Kata kunci: *DIF*, model Rasch, Mantel-Haenszel, *Chi-square* Scheuneman, teori tes klasik, Ujian Nasional

PENDAHULUAN

Dalam melakukan analisis suatu butir tes digunakan data yang berupa respon peserta tes untuk mengestimasi kemampuan peserta tes dan parameter butir. Peserta tes mendapatkan skor di atas kemampuan yang sebenarnya, atau sebaliknya mendapat skor yang jauh di bawah kemampuannya. Ketimpangan skor ini dapat disebabkan karena butir tes dan karena bukan butir tes. Ketimpangan yang disebabkan karena butir tes itulah yang disebut dengan bias butir. Ketimpangan skor yang telah dijelaskan di atas harus dihilangkan, diupayakan berkurang, atau paling tidak dideteksi besar dan arahnya. Hasil deteksi ini dapat digunakan untuk koreksi atau perbaikan butir-butir soal di masa

datang. Hambleton dan Roger mengemukakan pengertian bias butir yaitu individu-individu yang memiliki kemampuan sama dari kelompok berbeda mempunyai probabilitas yang tidak sama untuk menjawab dengan betul butir tersebut (Kartowagiran, 2004: 23).

Butir-butir yang berpotensi mengandung bias kemudian dianalisis dengan analisis logik seperti halnya mengapa butir ini relatif lebih sulit untuk satu kelompok. Apabila suatu butir relatif lebih sulit untuk suatu kelompok dan kesulitan tersebut tidak relevan terhadap konstruk tes maka butir tersebut dikatakan bias. Beberapa ahli mengganti istilah bias butir dengan nama *Differential Item Function (DIF)*. Istilah ini mencerminkan tujuan dari metode pendeteksian bias butir dalam mengidentifikasi butir-butir yang memiliki fungsi berbeda untuk kelompok peserta tes yang berbeda (Naga, 1992: 441).

Dalam melakukan pengukuran diperlukan perangkat tes yang valid dan reliabel, agar dapat memperoleh hasil pengukuran yang sesuai dengan apa yang hendak diukur. Untuk mengetahui kualitas suatu alat ukur maka perlu dilakukan uji psikometrik terhadap alat ukur tersebut. Para ahli psikometrika telah menetapkan kriteria bagi suatu alat ukur psikologis untuk dapat dinyatakan sebagai alat ukur yang baik dan mampu memberikan informasi yang tidak menyesatkan (Azwar, 1996: 68). Secara matematis bias butir dapat dinyatakan dalam bentuk probabilitas (Wardani, 2009: 123). Sebagai contoh, jika suatu butir soal secara sistematis lebih menguntungkan kelompok peserta wanita, butir soal tersebut mengandung bias yang positif terhadap wanita (bias gender), begitu pula sebaliknya.

Selain bias gender, ada pengelompokan lain seperti bias budaya, bias bahasa, dan bias etnik. Bias butir soal secara statistika dapat diestimasi arah dan besarnya sehingga dapat dilakukan koreksi secara statistika. *DIF* dapat diidentifikasi dan diukur dengan berbagai metode, salah satunya adalah melihat perbedaan probabilitas menjawab benar dari dua kelompok yang diteliti. Dengan kata lain, *DIF* adalah perbedaan probabilitas menjawab benar butir soal dari dua kelompok yang berbeda setelah mengontrol tingkat kemampuan (Crocker, 1986: 271). Bias butir memiliki dua karakter, yaitu arah (*direction*) dan besaran (*magnitude*). Besaran bias dapat diestimasi secara statistik.

Secara konseptual, *DIF* dikatakan muncul pada sebuah butir soal, jika peserta yang mempunyai kemampuan yang sama pada konstruk yang diukur oleh tes, tetapi dari kelompok yang berbeda, mempunyai peluang yang berbeda dalam menjawab benar soal tersebut (Hulin, 1983: 23). Konstruk yang sama misalnya mengukur hanya satu kemampuan atau *unidimensional* dan kelompok yang berbeda misalnya kelompok pria dan kelompok perempuan. Selanjutnya Hambleton (1991: 110) mengemukakan bahwa suatu butir menunjukkan *DIF* jika peserta tes memiliki kemampuan sama berada dalam kelompok yang berbeda, tidak mempunyai probabilitas sama untuk menjawab betul.

Umumnya terdapat dua jenis *DIF* (Lord, 1980: 114) yaitu 1) *DIF* uniform dan *DIF* tidak uniform. *DIF* uniform muncul jika keuntungan salah satu kelompok

terhadap kelompok lainnya terjadi pada setiap level kemampuan, dan 2) *DIF* tidak uniform muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya tidak terjadi pada setiap kemampuan. Jika dikaitkan dengan pengertian interaksi pada uji statistik analisis varian, *DIF* uniform terjadi jika tidak terdapat interaksi antara tingkat kemampuan peserta dan keanggotaan kelompok dan *DIF* tidak uniform terjadi jika terdapat interaksi antara tingkat kemampuan peserta dan keanggotaan kelompok (Hambleton, 1991: 105).

Untuk menentukan apakah suatu butir terindikasi suatu *DIF* atau tidak, diperlukan indeks *DIF*, yaitu indeks yang menunjukkan sekuat indikasi *DIF* ada pada butir soal itu. Dalam konteks teori responsi butir, terjadi atau tidaknya *DIF* pada sebuah butir soal terletak pada fungsi respons butir (*Item Response Function*) untuk butir tersebut pada kelompok yang dipersoalkan. Kurva yang menggambarkan fungsi respons disebut kurva respons butir atau kurva karakteristik. Untuk melakukan pendeteksian keberadaan *DIF* pada butir tes, sebuah populasi dibagi menjadi dua kelompok yaitu kelompok fokal dan kelompok referensi. Kelompok fokal merupakan kelompok yang diselidiki apakah ada butir yang mengandung *DIF* pada kelompok tersebut. Kelompok referensi merupakan kelompok pembanding.

Penelitian tentang berbagai metode pendeteksian untuk mendeteksi keberadaan *DIF* sudah banyak dilakukan. Kartowagiran (2004: 156) melakukan penelitian mengenai perbandingan sensitivitas berbagai metode dalam mendeteksi bias butir memberikan kesimpulan bahwa baik diuji dengan Friedman maupun dengan analisis faktor, ternyata metode *Likelihood Ratio Test* yang paling sensitif dalam mendeteksi bias butir diikuti dengan kurva karakteristik butir, *Chi-square* Lord, dan luasan Raju. Hasil penelitian tersebut sesuai dengan hasil penelitian Kim dan Cohen (1995: 304) yang membandingkan metode deteksi *Chi-square* Lord, luasan Raju, dan rasio kebolehdjian. Hasil penelitian menunjukkan bahwa ada kesesuaian yang baik dalam pendeteksian *DIF* di antara ketiga metode tersebut. Metode tes rasio kebolehdjian mampu mendeteksi *DIF* lebih banyak daripada metode *Chi-square* Lord, dan luasan Raju.

Susongko (2005: 18) melakukan penelitian mengenai penyeteraan parameter butir secara konkuren untuk menguji keberadaan *DIF*. Hasil penelitian menunjukkan bahwa hampir semua rayon yang melaksanakan ujian nasional di Jawa Tengah, butir-butir soalnya mengandung *DIF*. Ma'ruf (2004: 146) juga telah melakukan penelitian tentang bias jenis kelamin soal-soal olimpiade sains nasional bidang biologi pada tingkat provinsi. Hasil penelitian menunjukkan bahwa ada item soal yang mengandung bias berdasarkan jenis kelamin dari keseluruhan soal olimpiade sains bidang Biologi SLTP tingkat provinsi di DIY, ada empat butir yang mengarah atau menguntungkan kelompok peserta perempuan, dan ada satu butir yang mengarah atau menguntungkan kelompok pria. Hasil penelitian yang lain juga sudah dilakukan oleh Zulkarnaen (2010: 745) menunjukkan bahwa: Pertama, karakteristik perangkat tes berdasarkan teori tes klasik: parameter tingkat kesukaran ada 2 butir sukar, 27 butir sedang, dan 11

butir mudah. Jika ditinjau dari parameter daya pembedanya, maka ada 25 butir sangat baik, 11 butir baik, dan 3 butir cukup baik, dan 1 butir tidak baik. Dengan demikian berdasarkan teori tes klasik, ada 28 butir yang baik, dan dapat digunakan untuk pengembangan bank soal.

Hasil penelitian Rahayu (2010: 20) yang melakukan pendeteksian *DIF* dengan menggunakan metode Lord's *Chi-square test*. Hasil pendeteksian *DIF* melalui metode ini didasarkan pada banyaknya butir *false positive* yang mempengaruhi keakuratan metode *linking*. Jumlah butir *false positive* dihitung dari butir yang sebelumnya diperkirakan tidak *DIF* berdasarkan *likelihood ratio test*, akan tetapi setelah pendeteksian *DIF* menggunakan Lord's *Chi-square test*, ternyata ditemukan bahwa item mengandung *DIF*.

Tujuan penelitian ini adalah untuk memperoleh perbedaan antara metode *Chi-square* Scheuneman, metode Mantel Haenszel, dan metode teori responsi butir model Rasch yang paling sensitif dalam penelitian *DIF*.

METODE PENELITIAN

Penelitian ini merupakan eksperimen yang terdiri dari dua variabel. Variabel bebas dari penelitian ini adalah metode pendeteksian bias butir. Rancangan dari penelitian ini digambarkan dalam tabel 1 berikut:

Tabel 1. Rancangan Penelitian dengan Menggunakan Tiga Metode

Replikasi/ Pengulangan	Metode <i>Chi-square</i> Scheuneman	Metode Mantel -Haenszel	Metode TRB Model Rasch
1	$\chi^2_{\text{Scheuneman}}$	$\chi^2_{\text{Mantel-Haenszel}}$	$\chi^2_{\text{Raschmodel}}$
2	$\chi^2_{\text{Scheuneman}}$	$\chi^2_{\text{Mantel-Haenszel}}$	$\chi^2_{\text{Raschmodel}}$
3	$\chi^2_{\text{Scheuneman}}$	$\chi^2_{\text{Mantel-Haenszel}}$	$\chi^2_{\text{Raschmodel}}$
4	$\chi^2_{\text{Scheuneman}}$	$\chi^2_{\text{Mantel-Haenszel}}$	$\chi^2_{\text{Raschmodel}}$
5	$\chi^2_{\text{Scheuneman}}$	$\chi^2_{\text{Mantel-Haenszel}}$	$\chi^2_{\text{Raschmodel}}$
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
50	$\chi^2_{\text{Scheuneman}}$	$\chi^2_{\text{Mantel-Haenszel}}$	$\chi^2_{\text{Raschmodel}}$

Jenis data dalam penelitian ini adalah data sekunder yang berupa skor pekerjaan siswa. Skor hasil pekerjaan siswa berasal dari respons siswa terhadap perangkat tes Ujian Nasional (UN) SMA mata pelajaran matematika paket 13

untuk wilayah kota dan kabupaten Tangerang tahun 2009. Dalam penelitian ini sampel sumber data yang diambil sebanyak 5000 yang terdiri dari siswa pria sebanyak 2500 orang dan siswa perempuan sebanyak 2500 orang. Populasi butir adalah butir-butir yang terdapat dalam perangkat Ujian Nasional SMA mata pelajaran matematika. Banyaknya butir dalam perangkat tes adalah 40 butir.

Pendekatan *chi-square* Scheuneman. Pada dasarnya cara ini membagi skor peserta ke dalam sejumlah interval yang dinamakan interval skor atau interval kemampuan peserta. Populasi responden dibagi ke dalam subpopulasi yang diduga terkena bias butir, skor responden dibagi ke dalam interval. Butir tidak bias jika proporsi jawaban betul pada setiap interval adalah sama untuk dua subpopulasi itu. Pemeriksaan bias butir dengan persamaan berikut:

(a) statistik *chi-square* tiap interval adalah:

$$\chi_{Pk1}^2 = \frac{(A_{k1} - E_{Pk1})^2}{E_{Pk1}} \quad \chi_{Pk2}^2 = \frac{(A_{k2} - E_{Pk2})^2}{E_{Pk2}}, \text{ dan}$$

(b) *chi-square* Scheuneman pada K interval adalah:

$$\chi_s^2 = \sum_{k=1}^K \chi_{Pk1}^2 + \sum_{k=1}^K \chi_{Pk2}^2$$

Penggunaan metode Mantel-Haenszel berdasarkan asumsi-asumsi sebagai berikut: (1) hanya mengukur satu dimensi (unidimensi), (2) kemampuan peserta dinyatakan dengan skor total yang diperoleh peserta tes dari seluruh butir soal dengan menganggap setiap soal memiliki bobot yang sama, (3) level dari kemampuan peserta tes dapat digolongkan dalam M kelompok yang berurutan, dan (4) setiap peserta tes dapat dikelompokkan kepada satu dan hanya satu kelompok, yaitu kelompok acuan atau kelompok fokus (Budiyono, 2005: 67). Pendeteksian ada atau tidaknya *DIF* untuk perangkat soal Ujian Nasional diawali dengan mengelompokkan respons/jawaban siswa berdasarkan perbedaan gender, selanjutnya dilakukan analisis keberadaan *DIF* dengan metode Mantel-Haenszel dengan menggunakan persamaan sebagai berikut:

$$\alpha_{MH} = \frac{\sum_{k=1}^K \frac{M_{Rbk} M_{Fsk}}{M_k}}{\sum_{k=1}^K \frac{M_{Rsk} M_{Fbk}}{M_k}}$$

Perhitungan bias butir Mantel-Haenszel yang berdistribusi *chi-square* dilakukan dengan menggunakan rumus berikut:

$$\chi_{MH}^2 = \frac{\left[\left| \sum_{k=1}^K M_{Rbk} - \sum_{k=1}^K \frac{M_{bk} M_{Rk}}{M_k} \right| - 0,5 \right]^2}{\sum_{k=1}^K \frac{M_{bk} M_{sk} M_{Rk} M_{Fk}}{M_k}}$$

dengan distribusi *chi-square*. Ketentuan pengujian adalah jika $\chi_{MH}^2 \geq \chi_{tabel}^2$ maka item terindikasi *DIF*.

Pendekatan teori responsi butir model Rasch. Model Rasch yang digunakan mendeteksi keberadaan *DIF* dalam suatu butir soal pada penelitian ini merupakan model pengukuran yang didasarkan pada satu parameter yaitu tingkat kesukaran butir soal.

HASIL PENELITIAN

Dalam penelitian ini sebuah butir soal terindikasi *DIF* dengan metode Rasch apabila nilai *Chi-square* hitung lebih besar dari $\chi_{tabel}^2 = 14,067$ dengan taraf signifikansi 0,05 dan db= 7. Butir-butir yang terindikasi mengandung *DIF* sebanyak 17 butir soal, yaitu nomor: 1, 3, 4, 9, 10, 18, 19, 20, 21, 25, 28, 29, 30, 35, 36, 38, dan 40. Sebuah item terindikasi *DIF* dengan metode Mantel-Haenszel apabila Mantel-Haenszel *chi-square* suatu butir soal lebih besar dari $\chi_{tabel}^2 = 12,592$ dan db=6. Butir-butir yang terindikasi mengandung *DIF* dengan metode Mantel-Haenszel sebanyak 14 butir soal, yaitu nomor: 1, 9, 10, 12, 18, 19, 21, 26, 32, 36, 37, 38, 39, dan 40.

Sebuah butir soal terindikasi *DIF* dengan metode *chi-square* Scheuneman apabila nilai *chi-square* hitung lebih besar dari χ_{tabel}^2 . Butir-butir yang terindikasi mengandung *DIF* sebanyak 8 butir soal, yaitu nomor: 9, 19, 25, 34, 36, 38, 39, dan 40. Berdasarkan hasil pendeteksian *DIF* tersebut disimpulkan bahwa metode yang paling sensitif dalam mendeteksi keberadaan *DIF* adalah model Rasch sebanyak 17 butir soal, kemudian pada urutan kedua adalah dengan metode Mantel-Haenszel sebanyak 14 butir soal, sedangkan yang paling kurang sensitif dalam mendeteksi keberadaan adalah metode *chi-square* Scheuneman sebanyak 8 butir soal. Setelah dilakukan analisis butir soal yang terindikasi mengandung *DIF*, dapat ditunjukkan bahwa ada perbedaan jumlah butir yang terindikasi mengandung *DIF* jika dideteksi dengan pendekatan model Rasch, metode Mantel-Haenszel, dan metode *chi-square* Scheuneman sebagaimana tabel 2.

Hipotesis pertama dalam penelitian ini menyatakan bahwa metode TRB model Rasch lebih sensitif jika dibandingkan dengan Metode *chi-square* Scheuneman dalam menguji pendeteksian keberadaan *DIF*. Hipotesis pertama ini diuji dengan kriteria tolak H_0 jika $Z_{hitung} > Z_{tabel}$, hasil perhitungan menunjukkan $Z_{hitung} = 29,875 > Z_{tabel} = 2,330$ dan hasil perhitungan sehingga H_0 ditolak, yang berarti bahwa pendeteksian keberadaan *DIF* dengan metode TRB model Rasch lebih sensitif dibandingkan dengan metode *chi-square* Scheuneman.

Tabel 2. Rerata Jumlah Butir Terindikasi Mengandung *DIF* untuk Masing-masing Metode yang Digunakan dalam Penelitian.

Ket	Metode TRB Model Rasch		Metode Mantel-Haenszel		Metode <i>Chi-square</i> Scheuneman	
	T	TT	T	TT	T	TT
Σ	14,84	7,88	11,74	9,75	5,25	14,26
P	0,647	0,347	0,546	0,454	0,269	0,731

Keterangan : T = terindikasi *DIF*, TT = tidak terindikasi

Hipotesis kedua dalam penelitian ini menyatakan bahwa metode TRB model Rasch lebih sensitif jika dibandingkan dengan Metode Mantel-Haenszel dalam menguji pendeteksian keberadaan *DIF*. Hipotesis kedua ini diuji dengan kriteria tolak H_0 jika $Z_{hitung} > Z_{tabel}$, hasil perhitungan menunjukkan $Z_{hitung} = 7,949 > Z_{tabel} = 2,330$ sehingga H_0 ditolak, yang berarti bahwa pendeteksian keberadaan *DIF* metode TRB model Rasch lebih sensitif dibandingkan dengan metode Mantel-Haenszel. Hipotesis ketiga dalam penelitian ini menyatakan bahwa metode Mantel-Haenszel lebih sensitif jika dibandingkan dengan metode *chi-square* Scheuneman dalam menguji pendeteksian keberadaan *DIF*.

Hipotesis ketiga ini diuji dengan kriteria tolak H_0 , jika $Z_{hitung} > Z_{tabel}$ dan hasil perhitungan menunjukkan $Z_{hitung} = 23,014 > Z_{tabel} = 2,330$, sehingga H_0 ditolak yang berarti bahwa pendeteksian keberadaan *DIF* dengan metode Mantel-Haenszel lebih sensitif dibandingkan dengan metode *chi-square* Scheuneman. Berdasarkan hasil analisis deteksi *DIF* maka dapat disimpulkan bahwa metode teori response butir model Rasch merupakan metode yang paling sensitif dibandingkan dengan metode Mantel-Haenszel dan metode *chi-square* Scheuneman.

PEMBAHASAN

Berdasarkan hasil pendeteksian keberadaan *DIF*, ternyata metode deteksi *DIF* yang paling sensitif untuk tes Ujian Nasional Matematika SMA tahun 2009 adalah pendeteksian dengan metode TRB model Rasch diikuti dengan metode Mantel-Haenszel kemudian metode *chi-square* Scheuneman. Hasil penelitian ini sesuai dengan hasil penelitian Kim dan Cohen (1995: 296) yang membandingkan metode deteksi *chi-square* Lord, luasan Raju, dan tes rasio kebolehjadian. Hasil penelitian menunjukkan bahwa ada kesesuaian yang baik dalam pendeteksian *DIF* di antara ketiga metode itu. Metode tes rasio kebolehjadian mampu mendeteksi *DIF* lebih banyak daripada metode *chi-square* Lord, dan luasan Raju.

Butir-butir soal yang mengandung aspek spasial atau keruangan cenderung menguntungkan siswa pria, sedangkan butir-butir yang mengandung aspek verbal cenderung menguntungkan siswa wanita. Hal ini sejalan dengan

beberapa penelitian yang telah dilakukan menunjukkan bahwa kemampuan geometri kelompok pria lebih tinggi dibandingkan dengan kelompok wanita. Hal ini sejalan dengan hasil penelitian Linn dan Peterson pada tahun 1986; Stumpf dan Klieme pada tahun 1989; Masters dan Sander pada tahun 1993; Hedges dan Nowell pada tahun 1995; Voyer dan Bryden pada tahun 1995 juga menunjukkan bahwa rata-rata skor tes kemampuan spasial atau keruangan untuk pria lebih tinggi daripada skor wanita.

Selain itu juga, siswa wanita yang terbiasa mendapat perlakuan tidak adil dari gurunya dapat mengurangi rasa percaya dirinya, sehingga kemampuan geometri siswa kelompok wanita menjadi semakin rendah. Dunia pendidikan yang sudah berjalan berpuluh tahun ini, siswa wanita masih berjuang melawan ketidaksamaan gender yang substansial di sekolah. Dengan adanya ketidaksamaan ini wanita mengalami keterbatasan produktivitas sosial dan emosional. *The American Association of University Women* (Fakih, 2004: 34) menerbitkan laporan penting terkait dengan ketidaksamaan gender di dalam kelas. Laporan itu menjelaskan bahwa siswa wanita menerima perhatian dari guru yang lebih sedikit bila dibandingkan dengan siswa pria. Siswa wanita juga sering memperoleh kritikan pada pekerjaan yang telah diselesaikannya tanpa menghargai kebersihan, kerapian, dan ketepatan pengumpulan pekerjaan tersebut.

Penelitian yang telah dilakukan oleh Ridlo dan Azwar (2005: 5) juga menunjukkan hal yang sama. Konsep atau muatan materi dan pemicu *DIF* yang terkandung dalam butir-butir soal yang terdeteksi *DIF* yaitu: (1) muatan materi yang menguntungkan pria, adalah penerapan konsep diferensial pada luas dan keliling persegi panjang, pembacaan grafik fungsi sinus-cosinus, persamaan logaritma, peluang, dan persamaan elips, dan (2) muatan materi yang menguntungkan perempuan adalah persamaan fungsi sinus-cosinus, deret aritmatika, statistika (median dari data interval), nilai maksimum atau minimum pada suatu pertidaksamaan linier, penjumlahan dan perkalian vektor, proyeksi vektor, dan persamaan garis singgung lingkaran yang sejajar dengan persamaan garis tertentu.

Sumber pemicu yang diduga menimbulkan *DIF*, yaitu: (1) pemicu *DIF* yang menguntungkan pria adalah konteks, strategi *short cut*, penyelesaian bertahap tetapi tidak simultan, dan transformasi ke bentuk spasial, dan (2) pemicu *DIF* yang menguntungkan perempuan adalah hafalan, penyimbolan dan lambang, perhitungan berjenjang, manipulasi aljabar, perhitungan bertahap dan simultan, kesalahan ketik, dan transformasi ke bentuk spasial. Hasil penelitian ini, menunjukkan sebuah butir soal dapat mengandung *DIF* dikarenakan oleh konteks, muatan materi pada butir soal tersebut, atau bias juga karena strategi yang berbeda antara kelompok pria dan perempuan dalam menyelesaikan suatu permasalahan matematika. Secara tidak langsung memberikan suatu pemahaman bahwa, dari sudut pandang kelompok peserta pria dan perempuan,

perangkat tes Ujian Nasional Matematika SMA telah berfungsi secara relatif seimbang dengan sedikit menguntungkan kelompok peserta pria.

Kecenderungan siswa pria lebih diuntungkan pada butir dengan aspek keruangan tinggi dan siswa perempuan lebih diuntungkan pada butir dengan ketelitian tinggi itu betul, tetapi di tempat atau negara lain belum tentu. Hal ini dapat dipahami karena salah satu faktor yang mempengaruhi perbedaan perilaku wanita atau pria adalah *masculinity-femininity* yang dianut. Kata *masculinity* dipakai maka yang terbayang adalah sifat keras, kompetitif, dan pendekatan yang keras untuk membuat keputusan sehingga kadang-kadang mengabaikan perasaan yang terkena keputusan itu. Sementara itu dalam kata *femininity* mengandung keinginan untuk kooperatif, keinginan untuk menjadi pekerja yang menyenangkan, menciptakan kondisi kerja yang bagus dan lebih bijaksana dalam mengambil keputusan dengan memperhatikan perasaan orang yang dikenai keputusan.

Penelitian lain dengan menggunakan metode yang sama telah dilakukan oleh Budiyo (2009: 18) menyimpulkan sebagai berikut: (1) urutan ketepatan metode pendeteksian *DIF* dari tiga metode yang digunakan, mulai dari yang paling tepat, adalah metode Sibtest, metode Mantel-Haenszel, dan metode regresi logistik, (2) pada kandungan *DIF* paling banyak 8% metode Mantel-Haenszel tidak melakukan kesalahan, dan (3) metode Sibtest tidak melakukan kesalahan sama sekali pada setiap persentase kandungan *DIF*. Sejalan dengan hasil penelitian tersebut, maka Fidalgo, Ferreres, dan Muniz (2004: 925) juga telah meneliti tentang pendeteksian keberadaan *DIF* dengan menggunakan metode Mantel-Haenszel. Metode ketiga yang digunakan dalam penelitian ini adalah menggunakan metode *chi-square* Scheuneman. Hasil pendeteksian *DIF* dengan metode ini terdeteksi sebanyak 8 butir soal.

Hasil penelitian ini baru mengetahui metode deteksi *DIF* menggunakan data real, belum menggunakan simulasi untuk menyelidiki sifat-sifat *DIF*, sehingga dapat dilakukan penelitian terkait perbandingan sensitivitas metode deteksi *DIF* dengan simulasi yang melibatkan jumlah butir, banyaknya peserta, dan banyaknya dimensi. Penelitian ini menggunakan model logistik satu parameter, sehingga perlu dilakukan penelitian lanjutan untuk model logistik dua dan tiga parameter untuk menguji keakuratan ketiga metode tersebut. Selain itu juga perlu dilakukan penelitian pada teori responsi butir multidimensi (*multidimensional item response theory*) yang terdiri dari dua model, yaitu *compensatory* dan *noncompensatory*. Penelitian pendeteksian *DIF* sulit dilakukan, karena banyaknya tahap yang harus dilalui, sehingga perlu dikembangkan program atau *software (batch-file)* untuk memerintah komputer agar beberapa program dapat berjalan otomatis dan secara otomatis pula membaca hasil suatu output agar menjadi input program yang lainnya.

SIMPULAN

Pendeteksian dengan TRB model Rasch ternyata lebih sensitif dalam mendeteksi keberadaan *DIF* daripada metode Mantel-Haenszel dan metode *chi-square* Scheuneman, dan metode Mantel-Haenszel ternyata lebih sensitif dalam mendeteksi keberadaan *DIF* daripada dengan menggunakan metode *chi-square* Scheuneman. Butir-butir soal Ujian Nasional Matematika yang mengandung aspek spasial atau keruangan cenderung lebih menguntungkan siswa kelompok pria, sedangkan butir-butir soal yang mengandung aspek verbal atau kata-kata lebih cenderung menguntungkan siswa kelompok perempuan.

DAFTAR PUSTAKA

- Azwar, S. (1996). *Dasar-Dasar Psikometri*. Yogyakarta: Pustaka Pelajar.
- Budiyono. (2005). "Perbandingan Metode Mantel-Haenszel, SIBTEST, Regresi Logistik dan Perbedaan Peluang dalam Mendeteksi Keberadaan DIF." *Disertasi*. Yogyakarta: Universitas Negeri Yogyakarta.
- Crocker, L., dan James A. Lagina. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart, and Winston.
- Fidalgo, Angel M., Doris Ferreres, dan Jose Muniz. (2004). "Utility of The Mantel-Haenszel Procedure for Detecting Differential Item Functioning in Small Samples." *Educational and Psychological Measurement*, Vol. 64.
- Lord, Frederic M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New Jersey: Lawrence Erlbaum Associates Publishers.
- Hedges, L. V., dan A. Nowell. (1995). "Differences In Mental Test Scores, Variability, And Numbers Of High-Scoring Individuals." *Science*. Vol. 269.
- Hambleton, Ronald. K., H. Swaminathan, dan H. Jane Rogers. (1991). *Fundamentals of Item Response Theory*. California: Sage Publications.
- Hulin, Charles L., Frits Drasgow, dan Charles K. Parsons. (1983). *Item Response Theory Applications to Psychological Measurement*. Illinois: Dow Jones-Irwin.
- Fakih, Mansour. (2004). *Analisis Gender dan Transformasi Sosial*. Yogyakarta: Pustaka Pelajar.
- Fraenkel, Jack R., dan Norman E. Wallen. (1990). *How to Design and Evaluate Research in Education*. New York: McGraw-Hill.

- Kartowagiran, B. (2004). "Perbandingan Berbagai Metode untuk Mendeteksi Bias Butir." *Disertasi.*, Yogyakarta: Universitas Gadjah Mada.
- Kim, Seock-Ho dan Allan S. Cohen, (1995). "A Comparison of Lord's Chi square, Raju's Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Functioning." *Journal of Applied Measurement in Education*. Vol. 8(2).
- Linn, Marcia C., dan Anne C. Peteresen. (1985). "Emergency and Characterization of Sex Differences In Spatial Ability: A Meta Analysis." *JSTOR Child Development*, Vol. 56(6).
- Masters, M. S., dan B. Sanders. (1993). "Is The Gender Difference In Mental Rotation Disappearing?". *Behavior Genetics*, Vol. 23.
- Ma'ruf, Ahmad. (2004). "Implementasi Teori Responsi Butir dalam Mendeteksi Keberadaan DIF." *Jurnal Pendidikan dan Kebudayaan*, Vol. 10(4).
- Naga, Dali S. (1992). *Pengantar Teori Skor Pada Pengukuran Pendidikan*. Jakarta: Universitas Gunadarma.
- Rahayu, Wardani. (2010). "Metode *Linking Butir False Positive* pada Pendeteksian DIF Berdasarkan Teori Responsi Butir." *Jurnal Penelitian dan Evaluasi Pendidikan*, Tahun 14, No. 1.
- Ridlo, A., dan Saefudin Azwar. (2005). "Keberfungsian Aitem Tes UAN Matematika SMA di Propinsi DIY Tahun Pelajaran 2003/2004." *Disampaikan pada Seminar Nasional: Hasil Penelitian tentang Evaluasi Hasil Belajar serta Pengelolaannya*. Pascasarjana UNY didukung oleh Direktorat P2TK & KPT dan HEPI, Yogyakarta.
- Stumpf, Heinrich dan Ekhard Klieme. (1989). "Sex-related differences inspatial ability: More evidence for convergence." *Perceptual and Motor Skills*, Vol. 69.
- Susongko, Purwo. (2005). "Penyetaraan Parameter Butir Secara Konkuren untuk Menguji Secara Statistik Keberadaan Item Function (DIF)." *Makalah disampaikan pada Seminar Nasional: "Hasil Penelitian tentang Evaluasi Hasil Belajar serta Pengelolaannya"*, Pascasarjana UNY didukung oleh Direktorat P2TK & KPT dan HEPI, Yogyakarta.
- Voyer, D., Voyer, S., & Bryden, M. P. (1995). "Magnitude Of Sex Differences In Spatial Abilities: A Meta-Analysis And Consideration Of Critical Variables." *Psychological Bulletin*. Vol. 117.

Wardani, Yulia Nugaan. (2009). "Perbedaan Sensitifitas Metode Analisis Faktor Konfirmatori (AFK) dan Model Persamaan Struktural." *Jurnal Pendidikan dan Kebudayaan*. Vol. 15(1).

Zulkarnaen. (2010). "Perbandingan Sensitivitas antara Metode Mantel-Haenszel dan Pendekatan Model Rasch dalam Mendeteksi DIF." *Jurnal Pendidikan dan Kebudayaan*, Vol. 17(2).