PERBANDINGAN METODE *MANTEL-HAENSZEL, SIBTEST*, REGRESI LOGISTIK, DAN PERBEDAAN PELUANG DALAM MENDETEKSI KEBERBEDAAN FUNGSI BUTIR

Oleh:
Budiono

Abstrak

Penelitian ini bertujuan untuk mengetahui urutan sensitivitas dan ketepatan deteksi *DIF* antara metode *Mantel-Haenszel*, regresi logistik, SIBTEST, dan perbedaan peluang, baik secara keseluruhan maupun jika ditinjau dari distribusi kemampuan peserta tes, ukuran sampel, dan panjang tes.

Penelitian ini terdiri dari penelitian simulasi dan dengan data riil. Data simulasi dikembangkan dengan menggunakan model logistik tiga parameter. Data riil, diambil secara random 600 siswa laki-laki sebagai kelompok acuan dan 600 siswa perempuan sebagai kelompok fokus dari siswa SMA/MA jurusan IPA yang menempuh UAN Matematika tahun ajaran 2003/2004 di Kota Surakarta.

Studi simulasi menyimpulkan hal-hal berikut: (1) urutan sensitivitas metode, mulai dari yang paling sensitif: (a) regresi logistik, (b) *Mantel-Haenszel*, dan (c) perbedaan peluang atau SIBTEST; (2) pada tiap-tiap kategori distribusi kemampuan peserta tes, urutan sensitivitas metode, mulai dari yang paling sensitif: (a) regresi logistik, (b) *Mantel-Haenszel*, dan (c) perbedaan peluang atau SIBTEST; (3) pada tiap-tiap kategori ukuran sampel dan panjang tes, urutan sensitivitas metode, mulai dari yang paling sensitif: (a) regresi logistik atau *Mantel-Haenszel* dan (b) perbedaan peluang atau SIBTEST; (4) urutan ketepatan metode, mulai dari yang paling tepat adalah: (a) SIBTEST, (b) perbedaan peluang, dan (c) regresi logistik atau *Mantel-Haenszel*. 5) pada tiap-tiap kategori distribusi kemampuan peserta tes, ukuran sampel, dan panjang tes, metode yang mempunyai ketepatan paling tinggi selalu sama, yaitu metode SIBTEST disusul oleh metode perbedaan peluang.

Kata kunci: *mantel-haenszel, sibtest, regresi logistik, perbedaan peluang, mendeteksi keberbedaan fungsi butir.*

Pendahuluan

Suatu butir soal disebut baik apabila adil terhadap peserta tes yang mempunyai kemampuan yang sama walau berasal dari kelompok yang berbeda. Pengujian untuk melihat apakah butir soal bertindak adil atau tidak, disebut pengujian keberbedaan fungsi butir (*differential item functioning*), yang untuk selanjutnya disingkat *DIF*.

Secara konseptual, *DIF* muncul pada sebuah butir soal, jika peserta tes yang mempunyai kemampuan yang sama pada konstraks yang diukur oleh tes, tetapi berasal dari kelompok berbeda mempunyai peluang berbeda dalam menjawab benar butir soal tersebut (Hulin, Drasgow & Parson, 1993: 152, Roussos, Schnipke & Pashley, 1999: 293; Penfield & Lam, 2000: 6). Dalam konteks teori tes modern, yang terkenal dengan nama teori respons butir, terjadi atau tidak terjadinya *DIF* pada sebuah butir soal terletak kepada fungsi respons butir untuk butir soal tersebut pada kelompok yang dipersoalkan. Kurva yang menggambarkan fungsi respons butir disebut kurva respons butir atau kurva karakteristik butir (*item characteristic curve/ICC*). Jika sebuah butir soal mempunyai fungsi respons butir yang tepat sama untuk setiap kelompok, peserta tes pada setiap tingkat kemampuan q mempunyai peluang yang tepat sama untuk menjawab benar, terlepas dari keanggotaan kelompok. Butir soal yang demikian merupakan butir soal yang tidak memuat *DIF*. Sebaliknya, jika sebuah butir soal mempunyai fungsi respons butir yang berbeda untuk kelompok yang berbeda, butir soal tersebut memuat atau terkena *DIF*.

Terdapat dua jenis *DIF*, yaitu *DIF* uniform dan *DIF* non-uniform. *DIF* uniform muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya terjadi pada setiap level kemampuan, sedangkan *DIF* non-uniform muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya tidak terjadi pada setiap level kemampuan (Penfield & Lam 2000: 9).

Penyebab munculnya *DIF*, yang menunjukkan bahwa suatu butir soal tersebut berfungsi berbeda, dapat bermacam-macam, antara lain karena butir soal tersebut menguntungkan salah satu kelompok karena susunan bahasanya atau karena substansi yang ditanyakan lebih dikenal oleh salah satu kelompok. Penyebab munculnya *DIF* dapat juga karena adanya perbedaan fasilitas

antarkelompok, adanya perbedaan kemampuan guru yang mengajar, dan adanya pelaksanaan tes yang tidak adil.

Untuk butir-butir soal yang terkena *DIF* harus dilakukan pembahas-an lebih lanjut apakah butir-butir soal tersebut akan tetap dipakai atau dibuang dari sebuah tes. Jika penyebab munculnya *DIF* tidak terkait dengan konstraks yang diukur oleh tes, misalnya karena menggunakan istilah yang lebih dikenal oleh suatu kelompok dibandingkan dengan kelompok lain, maka butir soal yang terkena *DIF* tersebut harus dibuang dari tes. Jika tidak demikian halnya, butir soal tersebut harus dipertahankan dari tes, namun tetap diperlukan langkah-langkah lanjutan untuk menghilangkan sumber adanya *DIF* tersebut.

Identifikasi mengenai ada atau tidaknya *DIF* pada suatu tes merupakan suatu kebutuhan karena ketiadaan *DIF* dalam suatu tes dianggap sebagai aspek penting dari keadilan dalam pelaksanaan pengujian (Rudas & Zwick, 1997: 31; French & Miller, 1996: 315). Dengan kata lain, pen-deteksian *DIF* memainkan peran penting untuk memperoleh keadilan antarsub-populasi (Bolt, 2000: 307). Berpegang dari pendapat tersebut, suatu tes yang tidak bebas dari *DIF* dapat merugikan peserta tes. Oleh karena itu, pendeteksian *DIF* harus menjadi bagian esensial dari pengem-bangan tes. Namun demikian, jika karena sesuatu dan lain hal, pendetek-sian *DIF* tidak dapat dilakukan pada tahap pengembangan tes, pendetek-sian *DIF* dapat dilakukan pada saat setelah tes selesai dilaksanakan (Gierl, Khaliq & Boughton, 1999: 16). Hasil pendeteksian *DIF* yang dilakukan pada atau setelah pelaksanaan tes dapat dipakai sebagai umpan balik pengembangan tes pada masa-masa berikutnya atau untuk keperluan lain.

Pada pelaksanaan pendeteksian *DIF*, kelompok yang diselidiki apakah ada butir yang bias padanya disebut kelompok fokus (*focal group*) dan kelompok pembandingnya disebut kelompok acuan (*reference group*). Pengelompokan itu dapat berdasarkan umur, gender, ras atau etnis, kultur, kecacatan, atau kelompok kebahasaan. Dalam perspektif gender, misalnya, kelompok perempuan dapat ditentukan sebagai kelompok fokus dan kelompok acuannya adalah kelompok laki-laki; atau sebaliknya kelompok laki-laki ditentukan sebagai kelompok fokus dan kelompok acuannya adalah kelompok perempuan.

Di Indonesia, beberapa peneliti telah mendeteksi keberadaan *DIF* pada beberapa ujian atau tes, misalnya Purwo Susongka (2000), Heri Retnawati (2003), dan Badrun Kartowagiran (2005). Purwo Susongko (2000) meneliti keberadaan *DIF* pada EBTANAS (Evaluasi Belajar Tahap Akhir Nasional) untuk mata uji Kimia tahun 1999 di Provinsi Jawa Tengah. Dengan pengelompokan berdasarkan rayon kota dan nonkota, dapat disimpulkan bahwa terdapat sejumlah besar (6 – 12 buah) butir soal yang terkena *DIF*. Heri Retnawati (2003) mendeteksi keberadaan *DIF* pada tes masuk di tiga SMP di Yogyakarta pada tahun 2002 untuk mata uji Matematika dengan pengelompokan berdasarkan jenis kelamin. Hasil penelitiannya menunjukkan bahwa: (a) pada tes masuk SMP pertama, terdapat 2 butir soal yang terkena *DIF*; (b) pada tes masuk SMP kedua, terdapat 3 butir soal yang terkena *DIF*; dan (c) pada tes masuk SMP ketiga, terdapat 6 butir soal yang terkena *DIF*. Badrun Kartowagiran (2005) mendeteksi keberadaan *DIF* pada UAN Matematika SMP tahun 2003 di Yogyakarta dengan pengelom-pokan berdasarkan jenis kelamin. Hasil penelitiannya menyimpulkan terdapat 10 butir soal yang terkena *DIF*.

Temuan Purwo S, Heri R, dan Badrun Kw tersebut memberikan indikasi mengenai masih adanya *DIF* pada soal-soal ujian penting di Indonesia. Mengingat adanya *DIF* pada suatu ujian dapat menimbulkan ketidakadilan bagi peserta tes, selanjutnya para pengembang ujian di Indonesia harus melakukan pendeteksian *DIF* pada butir-butir soal yang diproduksinya. Pada masa sekarang, pngembangan tes yang berkualitas baik tidak cukup hanya dilihat dari sisi fisik, substansi, validitas, reliabilitas, daya pembeda, tingkat kesulitan, dan berfungsinya pengecoh saja, tetapi harus dilihat juga dari sisi ada atau tidaknya *DIF*.

Dalam pada itu, banyak metode pendeteksian DIF yang dikemuka-kan oleh para pakar, antara lain: metode *Mantel-Haenzel*, metode SIBTEST, metode regresi logistik, dan metode perbedaan peluang. Ketiga metode pertama berdasarkan teori tes klasik, sedangkan metode keempat berdasarkan teori respons butir.

Pada tahun 1959, Mantel dan Haenszel menampilkan prosedur untuk suatu studi pemadanan kelompok, yang oleh Holland dan Thayer (1988: 129) dipakai untuk mendeteksi *DIF*, yang kemudian terkenal dengan metode *Mantel-Haenszel*. Metode ini digunakan di *Educational Testing Service* di Amerika Serikat (Dorans & Holland, 1993: 38). Pada penggunaan metode *Mantel-Haenszel*, peserta tes pada tiap-tiap kelompok (fokus dan acuan) digolongkan menjadi M buah kategori berdasarkan pada level kemampuan peserta tes. Kemampuan peserta tes ini disebut variabel pemadanan (*matching variable*) (Holland & Thayer, 1993: 39). Kemampuan peserta tes tersebut diwakili oleh skor total peserta tes. Pada metode ini, indeks DIF dinyatakan oleh persamaan berikut (Holland & Thayer, 1988: 134; Dorans & Holland, 1993: 40):

$$\hat{\alpha}_{MH} = \frac{\sum_m \frac{R_{rm} W_{fm}}{N_{tm}}}{\sum_m \frac{R_{fm} W_{rm}}{N_{tm}}} \qquad (1)$$

dengan R_{fm} adalah banyaknya peserta kelompok fokus yang menjawab benar, W_{fm} adalah banyaknya peserta kelompok fokus yang menjawab salah, R_{rm} adalah banyaknya peserta kelompok acuan yang menjawab benar, W_{rm} adalah banyaknya peserta kelompok acuan yang menjawab salah, dan N_{tm} adalah banyaknya seluruh peserta tes pada level kemampuan m. Jika $\hat{\alpha}_{MH} > 1$, maka butir yang diselidiki menguntungkan kelompok acuan. Jika $\hat{\alpha}_{MH} < 1$, butir yang diselidiki menguntungkan kelompok fokus. Untuk menguji signifikansi, digunakan uji khi-kuadrat yang mempunyai derajad kebebasan 1 sebagai berikut (Holland & Thayer, 1988: 134; Dorans & Holland, 1993: 40):

$$MH_{\chi^2} = \frac{\left[\left|\sum_m R_{rm} - \sum_m E(R_{rm})\right| - 0{,}5\right]^2}{\sum_m Var(R_{rm})} \qquad (2)$$

dengan $E(R_{rm}) = \frac{N_{rm}R_{tm}}{N_{tm}}$ dan $Var(R_{rm}) = \frac{N_{rm}R_{tm}N_{fm}W_{tm}}{N_{tm}^2(N_{tm}-1)}$.

SIBTEST (*the Simultaneous Item Bias Test*) pertama kali dikemukan oleh Shealy dan Stout (1993: 198). Metode ini dapat digunakan untuk mendetek-si *DIF* pada tingkat butir atau pada tingkat *testlet* (kumpulan butir). Namun, pada penelitian ini hanya diteliti adanya *DIF* pada tingkat butir. Dengan metode SIBTEST, keseluruhan butir soal dibagi menjadi dua subtes, yaitu *the studied subtest* atau *the suspect subtest* dan *the matching subtest. The studied subtest* berisi butir soal yang diselidiki *DIF*-nya, yang untuk selanjutnya disebut subtes yang diselidiki. *The matching subtest* berisi butir-butir soal sisanya, yang untuk selanjutnya disebut subtes pemadanan. Subtes pe-madanan dipakai untuk mengelompokkan kedua kelompok (yaitu kelompok fokus dan kelompok acuan) ke dalam sub-subkelompok yang komparabel pada kemampuan yang diukur. Indeks *DIF* untuk metode SIBTEST diberikan oleh (Hsin-Hung Li, Nandakumar & Stout, 1995: 5):

$$\hat{\beta}_U = \sum_{k=0}^{K} \hat{p}_k(\bar{Y}^*_{Rk} - \bar{Y}^*_{Fk}) \qquad (3)$$

dengan $\hat{p}_k$ = proporsi peserta tes pada kelompok acuan dan kelompok fokus pada subkelompok X = k; $\bar{Y}^*_{Rk}$ = rerata tersesuaikan (*adjusted mean*) untuk kelompok acuan pada skor subtes pemadanan X = k; dan $\bar{Y}^*_{Fk}$ = rerata tersesuaikan (*adjusted mean*) untuk kelompok fokus pada skor subtes pemadanan X = k. Nilai dan pada persamaan di atas dicari dari suatu formula tertentu (Hsin-Hung Li, Nandakumar, & Stout, 1995: 8). Jika > 0 dan signifikan, butir soal yang bersangkutan terdeteksi *DIF* yang menguntungkan kelompok acuan. Sebaliknya, jika < 0 dan signifikan, butir soal yang bersangkutan terdeteksi *DIF* yang menguntungkan kelompok fokus. Uji statistik untuk metode ini diberikan oleh formula berikut (Hsin-Hung Li, Nandakumar, & Stout, 1995: 6):

$$B = \frac{\hat{\beta}U}{\hat{\sigma}(\hat{\beta}U)} \quad (4)$$

yang berdistribusi normal baku dengan adalah estimasi kesalahan baku.

Penggunaan metode regresi logistik diperkenalkan pertama kali oleh Swaminathan dan Rogers (1990). Metode ini merupakan metode berdasarkan teori tes klasik yang juga populer di samping metode Mantel-Haenszel dan SIBTEST (Embretson & Reise, 2000: 251). Jika metode Mantel-Haenszel dan SIBTEST didesain untuk hanya mendeteksi *DIF* uniform, metode regresi logistik dapat dipakai untuk mendeteksi *DIF* uniform dan *DIF* tidak uniform sekaligus. Pada metode regresi logistik, peluang seseorang menjawab benar suatu butir soal mempunyai bentuk logistik berikut (Swaminathan & Rogers, 1990: 363):

$$P(u = 1) = \frac{e^z}{1 + e^z} \quad (5)$$

dengan menyatakan peluang peserta menjawab benar suatu butir soal tertentu. Pada metode ini, karena dicari perbedaan antarkelompok (yang menyatakan adanya *DIF* uniform) dan interaksi antara keanggotaan kelompok dan kemampuan peserta tes (yang menyatakan *DIF* tidak uniform), z dinyatakan dalam bentuk berikut (Camilli & Shepard, 1994: 125):

$$z = \delta + \tau_1 G + \tau_2 X + \tau_3 (GX) \quad (6)$$

dengan X adalah skor total yang diperoleh peserta tes dan adalah kelompok peserta tes, yang diberi kode 1 (untuk kelompok fokus) atau 2 (untuk kelompok acuan). Jika signifikan, terdapat *DIF* tidak uniform. Jika tidak signifikan, diperoleh: (1) jika signifikan dan > 0, maka butir yang diselidiki terdeteksi *DIF* yang menguntungkan kelompok acuan, (2) jika tidak signifikan, butir yang diselidiki tidak memuat *DIF*, dan (3) jika signifikan dan < 0, maka butir

yang diselidiki terdeteksi *DIF* yang menguntungkan kelompok fokus. Pada regresi logistik digunakan statistik uji khi-kuadrat Wald yang mempunyai derajat kebebasan 1, yang dirumuskan sebagai berikut (Field, 2000:181):

$$W = \frac{\tau_k^2}{[SE(\tau_k)]^2} \qquad (7)$$

Pada penggunaan metode perbedaan peluang, keseluruhan peserta tes juga dikelompokkan menjadi dua kelompok yang disebut kelompok acuan dan kelompok fokus. Namun demikian, jika estimasi parameter dilakukan terhadap tiap-tiap kelompok tersebut secara terpisah, estimasi yang diperoleh tidak terbandingkan sebab unit skala dari q pada tiap-tiap kelompok adalah sembarang. Oleh karena itu, diperlukan penyetaraan (*equating*) dulu untuk mendapatkan skala pada matriks yang sama. Salah satu metode penyetaraan adalah metode tes jangkar (*anchor test method*) (Camilli & Shepard, 1994: 63). Pada penyetaraan tersebut keseluruhan butir soal dikelompokkan menjadi dua bagian, yang disebut subtes yang diselidiki (*the studied subtest*) dan subtes jangkar (*the anchor subtest*). Dalam penelitian ini subtes yang diselidiki hanya berisi satu butir soal, yaitu butir soal yang dilihat apakah terkena *DIF* atau tidak, dan subtes jangkar adalah butir-butir soal sisanya dalam tes. Pada metode perbedaan peluang, perbedaan peluang, yaitu perbedaan dua *ICC*, di-definisikan sebagai berikut.

$$\Delta P_j = P_R(\theta_j) - P_F(\theta_j) \qquad (8)$$

Untuk mengestimasi *signed measure DIF*, didefinisikan indeks *DIF* sebagai berikut (Camilli & Shepard, 1994: 67):

$$SPD\text{-}\theta = \frac{\sum_{j=1}^{n_F} \Delta P_j}{n_F} \qquad (9)$$

Jika *SPD-q* bernilai positif, butir yang bersangkutan terindikasi *DIF* yang menguntungkan kelompok acuan. Sebaliknya, jika *SPD-q* bernilai negatif, butir yang bersangkutan terindikasi *DIF* yang menguntungkan kelompok fokus. Di samping didefinisikan *signed measure*, didefinisikan *unsigned measure DIF* yang dilambangkan dengan *UPD-q* sebagai berikut.

$$UPD\text{-}\theta = \frac{\sum_{j=1}^{n_F} |\Delta P_j|}{n_F} \qquad (10)$$

Jika indeks *UPD-q* tidak sama dengan nilai mutlak *SPD-q*, terjadi *DIF* tidak uniform. Sebaliknya, jika perbedaan antara *UPD-q* dan *SPD-q* kecil, tidak terjadi *DIF* tidak uniform. Untuk menguji signifikansi indeks *DIF* berdasarkan metode perbedaan peluang, digunakan pengukuran perbandingan model (*model comparison measures*) yang menggunakan tes rasio kebolehjadian (*likelihood ratio test*) (Camilli & Shepard, 1994:79).

Dalam pada itu, untuk dapat mendeteksi butir soal yang terkena *DIF* sebanyak mungkin, diperlukan metode pendeteksian *DIF* yang sensitif. Makin banyak dapat mendeteksi *DIF*, makin sensitif suatu metode. Kecuali unsur sensitivitas tersebut, metode yang baik harus dapat mendeteksi *DIF* secara tepat atau paling tidak dengan tingkat kesalahan yang kecil. Makin kecil tingkat kesalahan yang dilakukan, makin tepat suatu metode pendeteksian *DIF*. Oleh karena itu, suatu penelitian yang membandingkan sensitivitas dan ketepatan beberapa metode pendeteksian *DIF*, kemudian mencari urutannya sangat diperlukan.

Penelitian ini bertujuan untuk menentukan urutan sensitivitas dan urutan ketepatan antarmetode Mantel-Haenszel, SIBTEST, regresi logistik, dan perbedaan peluang dengan menggunakan simulasi, baik secara keseluruhan maupun jika ditinjau dari distribusi kemampuan peserta tes, ukuran sampel, dan panjang tes. Kecuali itu, penelitian ini juga ingin menemukan seberapa banyak butir *DIF* ada pada UAN Matematika SMA/MA Jurusan IPA tahun

2003/2004 di Kota Surakarta dan ingin melihat metode yang paling sensitif di antara keempat metode tersebut di atas pada hasil UAN tersebut.

Indikator sensitivitas suatu metode adalah banyaknya butir yang dapat dideteksi. Makin banyak suatu metode dapat mendeteksi *DIF*, makin sensitif metode tersebut. Dengan melakukan pembandingan sensitivitas antarmetode akan diperoleh urutan sensitivitas antarmetode.

Indikator ketepatan suatu metode adalah banyaknya butir yang terdeteksi DIF secara benar, banyaknya butir yang terkena kesalahan tipe I, dan banyaknya butir yang terkena kesalahan tipe II. Makin banyak suatu metode mendeteksi *DIF* secara benar, makin sedikit menghasilkan butir yang terkena kesalahan tipe I, dan makin sedikit menghasilkan butir yang terkena kesalahan tipe II, makin tepat metode tersebut mendeteksi DIF. Seperti halnya sensitivitas, dengan melakukan pembandingan ketepat-an antarmetode yang diteliti diperoleh urutan ketepatan antarmetode.

Metode Penelitian

1. Studi Simulasi

Kemampuan peserta tes ditunjukkan skor total yang diperoleh peserta tes. Pada penelitian ini, distribusi kemampuan kelompok acuan dan kelompok fokus dibedakan atas enam jenis, yaitu: (1) kelompok acuan dan kelompok fokus keduanya berdistribusi normal dan mempunyai variansi yang homogen, (2) kelompok acuan dan kelompok fokus keduanya berdistribusi normal dan mempunyai variansi yang tidak homogen, (3) kelompok acuan berdistribusi normal, kelompok fokus berdistribusi miring ke kanan, dan kedua kelompok mempunyai variansi yang homogen, (4) kelompok acuan berdistribusi normal, kelompok fokus berdistribusi miring ke kanan, dan kedua kelompok mempunyai variansi yang tidak homogen, (5) kelompok acuan berdistribusi miring ke kanan, kelompok fokus berdistribusi miring ke kanan, dan kedua kelompok mempunyai variansi yang homogen, dan (6) kelompok acuan berdistribusi miring ke kanan, kelompok fokus berdistribusi miring ke kanan, dan kedua kelompok mempunyai variansi yang tidak homogen. Distribusi tidak normal yang dipilih adalah distribusi miring ke kanan (*skewed to the right*).

Ukuran sampel pada penelitian ini dibedakan menjadi tiga kategori, yaitu: (1) tiap-tiap kelompok acuan dan fokus mempunyai ukuran 400; (2) tiap-tiap kelompok acuan dan fokus mempunyai ukuran 600; dan (3) tiap-tiap kelompok acuan dan fokus mempunyai ukuran 800.

Dilihat dari panjangnya, tes dibedakan atas dua macam, yaitu: (1) tes dengan panjang 20 butir (mewakili tes pendek); dan (2) tes dengan panjang 40 butir (mewakili tes panjang).

Dengan melihat macam distribusi kemampuan, ukuran sampel, dan panjang tes yang disebutkan di atas, terdapat 6 ´ 3 ´ 2 = 36 kasus data. Setiap satu kasus dipandang sebagai satu perangkat tes. Perangkat tes pada suatu kasus dipandang berbeda dengan perangkat tes pada kasus yang lain. Untuk memperkuat hasil penelitian, dilakukan 5 replikasi. Hal ini berarti bahwa ketika dilakukan pembandingan sensitivitas dan ketepatan metode maka pembandingan itu dilakukan pada 5 ´ 36 = 180 perangkat tes yang berbeda. Pada pembandingan sensitivitas dan ketepatan metode tersebut jika diperhatikan distribusi kemampuan peserta tes dilakukan dengan 30 perangkat tes, jika diperhatikan ukuran sampel dilakukan dengan 60 perangkat tes, dan jika diperhatikan panjang tes dilakukan dengan 90 perangkat tes.

Oleh karena penelitian dilakukan melalui studi simulasi, maka diperlukan pembuatan data (*data generation*). Data yang diperlukan adalah data mengenai respons peserta tes, yang berwujud nilai 1 (benar) atau 0 (salah), demikian hingga skor total peserta (yang menggambarkan kemampuan peserta) berdistribusi normal atau miring seperti yang disebutkan di atas, sebanyak peserta tes, dan dengan panjang tes yang ditentukan. Kecuali persyaratan-persyaratan tersebut, setiap butir soal harus memenuhi kelayakan sebagai butir soal yang baik. Persyaratan kelayakan tersebut ditinjau dari dua sisi, yaitu kelayakan dari teori sisi tes klasik dan kelayakan dari sisi teori respons butir.

Pemerolehan data untuk tiap-tiap kasus dan tiap-tiap kelompok dilakukan melalui sebuah program dengan bahasa pemrograman Fortran menggunakan Microsoft Fortran Power Station 4.0. Data dibangkitkan melalui pendekatan teori respons butir dengan tiga parameter, yaitu daya pembeda (a), tingkat kesulitan (b), dan *pseudo-guessing* (c). Nilai paramater a, b, dan c pada

pembangkitan data ini untuk selanjutnya diasumsikan sebagai nilai parameter yang sebenarnya ada pada populasi dan dipakai untuk menentukan adanya *true DIF* (*DIF* yang sebenarnya). Penghitungan kesalahan tipe I dan kesalahan tipe II suatu metode dilakukan dengan membandingkan butir yang terdeteksi *DIF* oleh suatu metode dengan *true DIF* tersebut.

Untuk mendeteksi *DIF* dengan menggunakan metode *Mantel-Haenszel*, dibuat program dengan menggunakan bahasa pemrograman Fortran. Program yang dikembangkan dapat mendeteksi semua butir soal pada suatu tes sekali jalan sehingga untuk 180 kasus diperlukan 180 kali eksekusi program. Tingkat signifikansi yang digunakan adalah á = 1%.

Seperti halnya pada metode *Mantel-Haenszel*, untuk mendeteksi *DIF* dengan menggunakan metode SIBTEST, dibuat program dengan meng-gunakan bahasa pemrograman Fortran. Program yang dikembangkan dapat mendeteksi semua butir soal pada suatu tes sekali jalan sehingga untuk 180 kasus diperlukan 180 kali eksekusi program. Tingkat signifikansi yang digunakan adalah á = 1%.

Untuk mendeteksi *DIF* dengan menggunakan metode regresi logistik, digunakan paket program statistik SPSS 10.0 *for Windows*. Pada metode regresi logistik ini, deteksi *DIF* hanya dapat dilakukan butir per butir sehingga untuk keseluruhan kasus diperlukan $5 \times ((18 \times 20) + (18 \times 40)) = 2400$ kali eksekusi program dengan SPSS. Tingkat signifikansi yang dipakai adalah á = 1%.

Untuk mendeteksi *DIF* dengan menggunakan metode perbedaan peluang, digunakan paket program BILOG Versi 3.07 untuk melakukan estimasi parameter semua butir soal yang diselidiki setelah dilakukan penyetaraan (*equating*) melalui metode tes jangkar (*anchor test method*) seperti yang dilakukan oleh Camilli dan Shepard (1994: 63). Nilai estimasi parameter tiap-tiap butir soal setelah penyetaraan dipakai sebagai masukan (*input*) dalam menentukan jenis *DIF* yang terdeteksi. Untuk menguji signifikansi pada metode perbedaan peluang dan penentuan jenis *DIF*-nya, dibuat program komputer dengan menggunakan bahasa pemrograman Fortran. Program yang dikembangkan dapat mendeteksi semua butir soal pada suatu tes sekali jalan sehingga untuk 180 kasus diperlukan 180 kali eksekusi program. Uji signifikansi menggunakan tingkat signifikansi á = 1%.

Seperti disebutkan di depan, nilai-nilai a, b, dan c pada saat pembangkitan data diasumsikan sebagai nilai a, b, dan c yang sebenarnya ada pada populasi untuk tiap-tiap butir soal yang bersangkutan. Nilai-nilai a, b, dan c ini dipakai untuk menentukan adanya *true DIF* pada setiap kasus. Penentuan jenis *true DIF* mengunakan teknik seperti pada penentuan jenis *DIF* pada metode perbedaan peluang, yaitu dengan menghitung terlebih dulu nilai *SPD-q* dan *UPD-q*, kemudian ditentukan jenis *DIF*-nya.

Untuk menentukan *true DIF*, dikembangkan sebuah program komputer dengan bahasa pemrograman Fortran. Program yang dikem-bangkan dapat menentukan adanya *true DIF* untuk seluruh butir soal pada suatu kasus sekali jalan. Dengan demikian, untuk menentukan adanya *true DIF* untuk seluruh kasus diperlukan 180 kali eksekusi program.

Untuk melihat ada atau tidaknya perbedaan sensitivitas antar-metode digunakan uji statistik analisis variansi dengan menggunakan paket program SPSS *for Windows Release 10* yang mengambil tingkat signifikansi á = 1%. Urutan sensitivitas antarmetode ditentukan dengan melihat urutan banyak-nya butir soal yang terdeteksi *DIF* dengan semaksimal mungkin meman-faatkan hasil analisis variansi yang telah dilakukan. Pada penelitian ini, urutan sensitivitas dimulai dari metode yang paling sensitif.

Perbedaan ketepatan antarmetode dilihat dari banyaknya butir yang terdeteksi *DIF* secara benar, banyaknya butir yang terkena kesalahan tipe I, dan banyaknya butir yang terkena kesalahan tipe II yang dibuat oleh tiap-tiap metode. Untuk melihat ada tidaknya perbedaan ketepatan, digunakan uji analisis variansi dengan kondisi yang sama seperti ketika melihat ada atau tidaknya perbedaan sensitivitas. Urutan ketepatan ditentukan oleh urutan banyaknya butir yang terdeteksi *DIF* secara benar, banyaknya butir yang terkena kesalahan tipe I, dan banyaknya butir yang terkena kesalahan tipe II, secara bersama-sama, dengan memanfaatkan semaksimal mungkin hasil analisis variansi seperti yang dilakukan ketika membicarakan urutan sensitivitas. Seperti halnya urutan sensitivitas pada penelitian ini, urutan ketepatan dimulai dari metode yang mempunyai ketepatan paling tinggi.

2. Data Riil

Data diambil dari Kantor Dinas Pendidikan dan Kebudayaan Provinsi Jawa Tengah. Berdasarkan data yang ada, secara random diambil 600 siswa laki-laki sebagai kelompok acuan dan 600 data siswa perempuan sebagai kelompok fokus dari data siswa yang mengikuti UAN Matematika SMA/MA jurusan IPA pada tahun ajaran 2003/2004 di Kota Surakarta.

Pemilihan pengelompokan berdasarkan jenis kelamin dilakukan dengan alasan pendeteksian *DIF* dengan pengelompokan berdasarkan jenis kelamin dapat dipakai untuk memahami adanya perbedaan kemampuan antara siswa laki-laki dan siswa perempuan dalam belajar matematika dan sebab-sebabnya. Hal semacam ini diperlukan untuk dapat menciptakan kesejajaran (*equity*) dalam pembelajaran matematika. Dalam konteks ini, *equity* diartikan "*females should learn exactly the same mathematics as do males, be able to perform the same on various measures of mathematical learning, and have the same personal feelings toward oneself and mathematics* (Fennema, 2000: 3).

Teknik pendeteksian *DIF* pada data riil, sama dengan teknik pendeteksian *DIF* pada studi simulasi. Sebelum dilakukan analisis *DIF*, semua butir UAN tersebut dilihat kelayakannya menurut teori tes klasik dan teori respons butir pada tiap-tiap kelompok acuan dan fokus seperti pada data simulasi. Hanya kepada butir-butir soal yang memenuhi kelayakan sajalah analisis *DIF* dilakukan. Urutan sensitivitas metode deteksi *DIF* pada data riil dilihat haya dari urutan banyaknya butir soal yang terdeteksi.

Hasil Penelitian dan Pembahasan

1. Studi Simulasi

a. Analisis Sensitivitas Metode Deteksi *DIF*

Berdadarkan banyaknya butir yang terdeteksi *DIF* sebagai indikator sensitivitas, diperoleh hal-hal berikut. Secara keseluruhan, urutan sensivitasnya adalah: (1) regresi logistik, (2) *Mantel-Haenszel*, dan (3) perbedaan peluang atau SIBTEST. Pada setiap klasifikasi distribusi kemampuan peserta tes, urutan sensitivitasnya adalah: (1) regresi logistik, (2) *Mantel-Haenszel*, dan (3) perbedaan peluang atau SIBTEST. Pada setiap kategori ukuran sampel, urutan

sensitivitasnya adalah: (1) regresi logistik atau *Mantel-Haenszel* dan (2) perbedaan peluang atau SIBTEST. Pada setiap kategori panjang tes, urutan sensitivitasnya adalah: (1) regresi logistik atau *Mantel-Haenszel* dan (2) perbedaan peluang atau SIBTEST.

Temuan penelitian simulasi ini sejalan dengan temuan penelitian Rogers dan Swaminathan (1993), yang menyimpulkan bahwa pada umum-nya metode regresi logistik lebih banyak mendeteksi butir *DIF* dibanding-kan dengan metode *Mantel-Haenszel.* Temuan penelitian simulasi ini juga sejalan dengan temuan Heri Retnawati (2003) yang menyimpulkan metode *Mantel-Haenszel* mendeteksi *DIF* lebih banyak dibandingkan dengan metode tes rasio kebolehjadian (yang pada penelitian ini disebut metode perbedaan peluang). Temuan penelitian ini juga sejalan dengan temuan Stoneberg (2004) yang menghasilkan kecenderungan bahwa metode *Mantel-Haenszel* mendeteksi lebih banyak dibandingkan dengan metode SIBTEST. Namun demikian, temuan penelitian simulasi ini agak bertentangan dengan temuan Gierl, Khaliq, dan Boughton (1999) yang menyatakan pada ujian Matematika, urutan sensitivitasnya adalah metode SIBTEST, regresi logistik, disusul *Mantel-Haenszel;* sedangkan pada ujian Sains, urutan sensitivitasnya adalah metode regresi logistik, SIBTEST, disusul *Mantel-Haenszel.* Berbedanya temuan penelitian simulasi dengan temuan peneliti-an Gierl, Khaliq, dan Boughton mungkin disebabkan karena pada kasus-kasus data simulasi termuat sejumlah besar butir *DIF,* sedangkan pada data yang dianalisis oleh Gierl, Khaliq, dan Boughton tidak.

Berdasarkan analisis data simulasi ditemukan bahwa distribusi kemampuan peserta tes, ukuran sampel, dan panjang tes tidak berpengaruh kepada urutan sensitivitas metode deteksi *DIF.* Temuan simulasi ini cocok dengan temuan Rogers dan Swaminathan (1993). Hal ini memberikan petunjuk bahwa distribusi skor total peserta tes, ukuran sampel, dan panjang tes, bukanlah faktor yang mempengaruhi urutan sensitivitas metode.

b. Analisis Ketepatan Metode Deteksi *DIF*

Berdasarkan analisis data mengenai ketepatan, diperoleh hal-hal berikut. Secara keseluruhan, urutan ketepatan metode adalah: (1) SIBTEST, (2)

perbedaan peluang, dan (3) regresi logistik atau *Mantel-Haenszel.* Distribusi peserta tes, ukuran sampel, dan panjang tes berpengaruh kepada urutan ketepatan metode dalam mendeteksi *DIF.* Namun, pada setiap keadaan, yang paling tepat adalah metode SIBTEST, disusul oleh metode perbedaan peluang.

Walaupun secara umum metode regresi logistik merupakan metode yang paling sensitif dibandingkan dengan metode yang lainnya, tetapi metode regresi logistik merupakan metode yang paling tidak tepat mendeteksi *DIF.* Sebaliknya, walaupun metode SIBTEST merupakan metode yang paling tidak sensitif, namun metode SIBTEST merupakan metode yang paling tepat mendeteksi *DIF.* Implikasi temuan ini adalah bahwa para pengguna harus mempertimbangkan metode mana yang akan dipakai untuk mendeteksi *DIF,* dengan mempertimbangkan dari dua sisi yang bertolak belakang, yaitu sensitivitas atau ketepatan.

Paling tepatnya metode SIBTEST dibandingkan dengan tiga metode yang lainnya, mungkin disebabkan oleh hal-hal berikut. Pertama, pada metode SIBTEST variabel pemadanannya berubah dari satu butir soal ke butir soal yang lain, sedangkan pada tiga metode yang lainnya variabel pemadanannya atau variabel jangkarnya tetap dari satu butir soal ke butir soal yang lain. Kedua, pada metode SIBTEST dilakukan penyesuaian rerata skor pada kelompok acuan dan kelompok fokus pada setiap level subtes pemadanan.

Hal lain yang menyebabkan metode *Mantel-Haenszel,* regresi logistik, dan perbedaan peluang tidak setepat metode SIBTEST dalam mendeteksi *DIF,* mungkin karena variabel pemadanan, pada metode *Mantel-Haenszel* dan regresi logistik, atau variabel jangkar, pada metode perbedaan peluang, terkontaminasi oleh butir-butir soal yang terkena *DIF.* Oleh karena itu, beberapa pakar pengukuran menganjurkan untuk melakukan purifikasi (*purification*) ketika memproses pendeteksian *DIF* (Holland & Thayer, 1988: 141; Camilli & Shepard, 1994: 94) karena makin tidak reliabel skor total sebagai variabel pemadanan, makin besar terjadi kesalahan tipe I (Penfield & Lam, 2000: 7). Holland dan Thayer menyarankan untuk melakukan purifikasi terhadap variabel pemadanan dengan cara tidak mengikutkan butir-butir yang terkena *DIF* ke dalam variabel pemadanan, kemudian mengulangi lagi proses penentuan *DIF*

dengan variabel pemadanan yang baru. Senada dengan itu, Camilli dan Shepard (1994: 94) menyarankan untuk melakukan purifikasi terhadap variabel jangkar (yaitu butir-butir soal yang dipakai sebagai jangkar dalam penyetaraan) dengan cara tidak mengikutkan butir-butir yang terkena *DIF* ke dalam variabel jangkar ketika melakukan penyetaraan, kemudian mengulangi lagi proses penentuan *DIF* dengan variabel jangkar yang baru. Dengan cara purifikasi diharapkan diperoleh deteksi *DIF* yang tepat.

Analisis terhadap data simulasi diketahui bahwa terdapat pengaruh distribusi kemampuan peserta tes, ukuran sampel, dan panjang tes terhadap urutan ketepatan metode dalam mendeteksi *DIF*. Namun demikian, pada berbagai situasi, metode SIBTEST dan metode perbedaan peluang mempunyai ketepatan yang lebih tinggi dibandingkan dengan metode *Mantel-Haenszel* dan metode regresi logistik. Pada situasi tertentu, metode *Mantel-Haenszel* lebih baik daripada metode regresi logistik, namun pada situasi yang lain dapat terjadi sebaliknya. Temuan ini memberikan implikasi praktis kepada para pengguna agar dalam memilih metode pendeteksian *DIF* sebaiknya memilih metode SIBTEST, jika hanya ingin mendeteksi *DIF* uniform saja, atau memilih metode perbedaan peluang, terutama jika ingin mendeteksi *DIF* uniform dan *DIF* tidak uniform sekaligus.

2. Data Riil Hasil UAN Matematika

Dari 40 butir soal yang ada, terdapat 26 butir soal yang layak, dan 14 butir soal yang tidak layak. Dari keempatbelas butir yang tidak layak, terdapat 7 butir soal yang terlalu mudah, 1 butir soal yang terlalu sukar, 6 butir soal yang daya pembedanya rendah, dan 2 butir soal yang tidak cocok (*fit*) dengan model logistik tiga parameter.

Seperti disampaikan di depan, hanya kepada butir-butir yang memenuhi kelayakan saja yang dianalisis *DIF*-nya. Setelah dianalisis, hasilnya sebagai berikut. Metode *Mantel-Haenszel* mendeteksi 2 butir soal, yaitu nomor 15 dan 35, metode SIBTEST mendeteksi 4 butir soal, yaitu nomor 4, 15, 26, dan 35, metode regresi logistik mendeteksi 2 butir soal, yaitu nomor 15, dan 35, dan metode perbedaan peluang mendeteksi 2 buah butir soal, yaitu nomor 15 dan

35. Berdasarkan hal tersebut, dapat disimpulkan bahwa metode SIBTEST mendeteksi butir *DIF* lebih banyak dibandingkan dengan ketiga metode yang lain.

Urutan sensitivitas yang diperoleh dari data riil bertentangan dengan urutan sensitivitas secara umum yang diperoleh dari data simulasi. Pada data riil, yang paling sensitif adalah metode SIBTEST, sedangkan pada data simulasi yang paling sensitif secara umum adalah metode regresi logistik. Pada keadaan seperti ini, kondisi data riil dapat dimaknai sebagai kejadian khusus pada data simulasi. Data simulasi pun pada kasus-kasus tertentu, metode SIBTEST juga mendeteksi lebih banyak dibandingkan dengan ketiga metode yang lain. Walaupun pada kasus-kasus tertentu, metode SIBTEST mendeteksi lebih banyak dibandingkan dengan ketiga metode lainnya. Namun, secara umum metode yang paling sensitif adalah metode regresi logistik, disusul oleh Mantel-Haenszel, perbedaan peluang dan SIBTEST.

Ketidakkonsistenan temuan penelitian data simulasi dan data riil mungkin disebabkan oleh banyaknya tes yang dianalisis. Pada data simulasi dianalisis 180 kasus (tes), sedangkan pada data riil hanya dianalisis sebuah tes.

Temuan data riil menunjukkan bahwa butir soal UAN Matematika nomor 4, nomor 15, dan nomor 26 terdeteksi *DIF* yang menguntungkan kelompok siswa laki-laki dan butir soal nomor 35 terdeteksi *DIF* yang cenderung menguntungkan kelompok siswa perempuan. Temuan tersebut menunjukkan bahwa ada pokok bahasan yang lebih dikuasai oleh siswa laki-laki dan ada pokok bahasan yang lebih dikuasai oleh siswa perempuan. Adanya perbedaan penguasaan tersebut tentu dipengaruhi oleh satu atau beberapa sebab. Berikut ini dibahas faktor-faktor yang *mungkin* menyebab-kan terjadinya *DIF* pada butir soal nomor 4, nomor 15, nomor 26, dan nomor 35. Diperlukan penelitian tersendiri untuk meyakinkan kebenaran dugaan tersebut.

Butir Soal Nomor 4:

Nilai $\sin 45^0 \cos 15^0 + \cos 45^0 \sin 15^0$ sama dengan ....

a. $1/2$ b. $1/2\sqrt{2}$ c. $1/2\sqrt{3}$ d. $1/2\sqrt{6}$ e. $1/2\sqrt{3}$

Kunci Jawaban: c

Terjadinya *DIF* yang menguntungkan kelompok laki-laki pada butir soal nomor 4 dapat disebabkan sebagian siswa perempuan menggunakan cara konvensional, seperti yang dikemukakan oleh Gallagher dan DeLisi (Fennema, et al, 1998: 4), yaitu dengan lebih dulu mencari nilai sin 45Ú, cos 15Ú, cos 45Ú, dan sin 15Ú, kemudian mengalikan dan menjumlahkannya seperti yang diminta oleh soal. Namun, cara ini tidak akan sampai kepada jawaban yang dikehendaki karena nilai cos 15Ú dan sin 15Ú tidak dapat dengan mudah dicari, kecuali jika memakai kalkulator. Di sisi lain, kelompok siswa laki-laki diduga menggunakan cara yang lebih cerdas, yaitu menggunakan rumus sin (á + â) = sin á cos â + cos á sin â. Jadi, terjadinya *DIF* pada butir soal nomor 4 mungkin disebabkan siswa laki-laki menggunakan cara yang lebih cerdas dibandingkan dengan siswa perempuan.

Butir Soal Nomor 15:

Dua dadu dilambungkan bersama-sama. Peluang muncul mata dadu pertama 3 dan mata dadu kedua 5 adalah ....

a. 6/36 b. 5/36 c. 4/36 d. 3/36 e. 1/36

Kunci Jawaban : **e**

Terdeteksinya butir soal nomor 15 sebagai butir *DIF* yang menguntungkan kelompok siswa laki-laki dapat disebabkan oleh faktor berikut. Butir soal yang kelihatannya sederhana tersebut merupakan butir soal yang cukup kompleks karena siswa harus dapat menentukan ruang sampel (*sample space*) pelambungan dua buah dadu bersama-sama. Siswa juga harus dapat menentukan kejadian pada eksperimen tersebut.

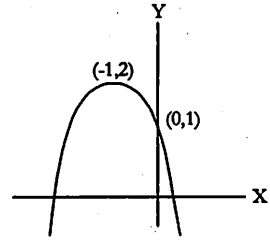
Kejadian pada eksperimen pada soal tersebut adalah K = {(3,5)} dengan ruang sampel S = {(1,1), (2,1), (3,1), ..., (4,6), (5,6), (6,6)}. Ada kemungkinan sebagian besar siswa perempuan tidak dapat menentukan kejadian dan ruang sampel pada eksperimen tersebut dengan benar karena tidak mengerti makna kejadian mata dadu pertama muncul 3 dan mata dadu kedua muncul 5. Ada kemungkinan, sebagian besar siswa perempuan menganggap kejadiannya adalah K = {3,5}, sehingga mereka tidak mendapatkan jawaban yang benar. Dengan demikian, terjadinya *DIF* pada butir soal nomor 15 dapat disebabkan adanya

kecenderungan bahwa siswa perempuan tidak dapat berpikir kompleks, seperti yang disebutkan oleh Fennema (2000: 8).

Butir Soal Nomor 26:

Persamaan grafik parabola pada gambar adalah ....

a. $y^2 - 4y + x + 5 = 0$
b. $y^2 - 4y + x + 3 = 0$
c. $x^2 + 2x + y + 1 = 0$
d. $x^2 + 2x - y + 1 = 0$
e. $x^2 + 2x + y - 1 = 0$



Kunci jawaban: **e**

Karena memuat grafik, butir soal nomor 26 mengandung unsur spasial. Untuk dapat mengerjakannya, siswa terlebih dulu harus mengetahui bentuk umum fungsi kuadrat yang mempunyai maksimum 2 di $x = -1$, kemudian melakukan substitusi untuk titik (0,1). Jadi, diperlukan analisis yang cukup panjang untuk dapat menjawabnya. Ada kemungkinan sebagian besar siswa perempuan tidak dapat mengerjakan butir soal ini karena tidak dapat melakukan analisis yang diperlukan. Jadi, terdeteksinya butir soal nomor 26 sebagai butir *DIF* yang menguntungkan kelompok laki-laki karena butir soal tersebut memuat unsur spasial (Lips, 1988: 126; Ricmond-Abbot, 1992: 48; Unger & Crawford, 1992: 78), juga diperlukan kemampuan kognitif tingkat tinggi untuk menyelesaikan butir soal tersebut (Fennema, 2000: 3).

Butir Soal Nomor 35:

Persamaan bayangan garis $3x - 2y + 1 = 0$ yang dicerminkan terhadap garis $y = x$ dilanjutkan rotasi sejauh [O,90°] adalah ....

a. $3x - 2y - 1 = 0$ c. $2x - 3y + 1 = 0$ e. $2x + 3y + 1 = 0$
b. $3x + 2y - 1 = 0$ d. $2x - 3y - 1 = 0$

Kunci jawaban: **b**

Untuk dapat mengerjakan butir soal tersebut, siswa harus dapat mengingat matriks pencerminan terhadap garis y = x, matriks rotasi [O,90°], kemudian dapat mencari bayangan garis 3x – 2y + 1 = 0 pada transformasi terakhir. Ada dua hal menonjol yang diperlukan dalam mengerjakan butir soal nomor 35, yaitu pengingatan kembali (*retrieval*) matriks transformasi dan ketelitian perhitungan dalam mengalikan matriks. Walaupun terkandung unsur spasial pada butir soal nomor 35, karena berkaitan dengan masalah geometri, unsur pengingatan kembali dan ketelitian perhitungan mungkin lebih menonjol. Oleh karena itu, butir soal nomor 35 terkena *DIF* yang menguntungkan kelompok perempuan, seperti yang dikatakan oleh Halpern (Stafslien, 2001: 2) dan Marshal (Carr, Jessup, & Fuller, 1999: 20).

Dari keempat butir soal UAN yang terkena *DIF*, tidak satu pun butir soal yang menguntungkan salah satu kelompok (siswa laki-laki atau siswa perempuan) karena susunan bahasa atau latar belakang budaya. Di sisi lain, baik siswa laki-laki maupun perempuan harus dapat menguasai pokok bahasan yang terkait dengan butir-butir soal tersebut karena tuntutan kompetensi minimal yang harus dikuasai. Oleh karena itu, keempat butir soal tersebut bukanlah butir soal yang bias sehingga harus tetap dipertahankan dalam tes. Adanya kandungan *DIF* pada keempat butir soal tersebut dapat dipakai oleh guru sebagai pijakan untuk menciptakan kesejajaran dalam pembelajaran matematika.

Simpulan

Berdasarkan penelitian simulasi yang telah dilakukan, dapat disimpulkan hal-hal berikut.

1. Secara keseluruhan, urutan sensitivitas metode deteksi *DIF* mulai dari yang paling sensitif, adalah: (1) regresi logistik, (2) *Mantel-Haenszel*, dan (3) perbedaan peluang atau SIBTEST.
2. Pada tiap-tiap kategori distribusi kemampuan peserta tes, urutan sensitivitas metode *DIF*, mulai dari yang paling sensitif, adalah: (1) regresi logistik, (2) *Mantel-Haenszel*, dan (3) perbedaan peluang atau SIBTEST.

3. Pada tiap-tiap kategori ukuran sampel dan panjang tes, urutan sensitivitas metode *DIF*, mulai dari yang paling sensitif, adalah: (1) regresi logistik atau *Mantel-Haenszel* dan (2) perbedaan peluang atau SIBTEST.
4. Secara keseluruhan, urutan ketepatan metode deteksi *DIF* mulai dari yang paling tepat, adalah: (1) SIBTEST, (2) perbedaan peluang, dan (3) regresi logistik atau *Mantel-Haenszel.*
5. Pada tiap-tiap kategori distribusi kemampuan peserta tes, ukuran sampel, dan panjang tes, metode yang mempunyai ketepatan paling tinggi adalah metode SIBTEST disusul oleh metode perbedaan peluang. Pada kasus-kasus tertentu metode *Mantel-Haenszel* lebih baik dibandingkan dengan metode regresi logistik, pada kasus-kasus tertentu metode *Mantel-Haenszel* lebih jelek dibandingkan dengan metode regresi logistik, dan pada kasus-kasus tertentu metode *Mantel-Haenszel* sama baiknya dengan metode regresi logistik.

Pada data riil hasil UAN Matematika pada Jurusan IPA di SMA/MA untuk Tahun Pelajaran 2003/2004 di Kota Surakarta diperoleh hal-hal berikut.

1. Dari 40 butir soal yang ada, hanya 26 butir soal yang memenuhi kelayakan dari sisi teori tes klasik dan dari sisi teori respons butir. Dari 26 butir soal yang memenuhi kelayakan, terdapat empat buah butir soal yang terkena *DIF.*
2. Metode SIBTEST merupakan metode yang paling sensitif dibanding-kan dengan ketiga metode lainnya. Metode SIBTEST mendeteksi *DIF* sebanyak 4 butir soal, sedangkan ketiga metode yang lainnya hanya mendeteksi 2 butir soal.

Saran-saran

Berdasarkan temuan penelitian dan pembahasan yang telah dilaku-kan, disarankan hal-hal berikut.

1. Kepada para pengguna metode deteksi *DIF*:
   a. Mengingat metode SIBTEST dan metode perbedaan peluang merupakan dua metode yang mempunyai tingkat ketepatan tinggi, jika tidak ada kendala pembuatan program atau telah tersedia paket programnya,

disarankan untuk menggunakan metode SIBTEST dan/atau metode perbedaan peluang dalam mendeteksi *DIF*.

b. Untuk meminimumkan kesalahan, disarankan untuk menggunakan dua atau tiga metode sekaligus, kemudian mengambil butir-butir yang terdeteksi oleh kedua atau ketiga metode tersebut sebagai butir *DIF* yang perlu dilakukan pembahasan lebih lanjut.

2. Kepada para pengembang tes:
   c. Dengan ditemukannya butir-butir soal yang memuat *DIF* pada UAN Matematika untuk SMA/MA Jurusan IPA Tahun Ajaran 2003/2004 di Kota Surakarta, kepada para pengembang tes, baik di pusat maupun daerah, khususnya untuk Matematika, disarankan agar melakukan prosedur sistematis untuk mendeteksi kemungkin-an adanya *DIF* sebelum suatu tes digunakan dan melakukan pembahasan untuk menentukan apakah butir-butir soal yang terindikasi *DIF* tersebut dipakai atau dibuang dari tes.
   d. Pada praksis pengembangan tes saat ini, terutama yang dilakukan di daerah, karena keterbatasan waktu atau karena sebab lain, pengembang tes tidak melakukan kajian terhadap butir-butir soal yang dikembangkan sebelum tes dilaksanakan. Namun demikian, pada umumnya mereka sudah menerapkan sistem komputerisasi dalam memeriksa hasil tes. Mengingat pentingnya pemenuhan prinsip keadilan bagi para peserta tes, kepada para pengembang tes disarankan untuk tetap melakukan kajian mengenai adanya *DIF* walaupun kajian tersebut dilakukan pada saat atau setelah suatu tes dilaksanakan. Para pengembang tes dapat membuat sebuah program yang dapat dipakai untuk memeriksa adanya *DIF* pada soal-soal yang dihasilkan bersamaan dengan waktu pemeriksaan hasil tes. Tentu saja, program tersebut dapat pula didesain untuk menilai kelayakan parameter butir-butir soal. Hasil kajian digunakan sebagai refleksi dan umpan balik bagi pengembangan tes pada tahun-tahun berikutnya.

3. Kepada para peneliti di bidang pengukuran dan pengujian:
   e. Terkait dengan adanya kesalahan dalam mendeteksi *DIF*, disaran-kan

menindaklanjuti penelitian ini untuk meneliti seberapa besar persentase butir *true DIF* ada pada satu perangkat tes agar suatu metode dapat bekerja dengan baik, dalam arti dapat memperoleh kesalahan sekecil mungkin.

f. Terkait dengan adanya ketidakcocokan antara hasil penelitian simulasi dan hasil penelitian yang menggunakan data riil dalam hal urutan sensitivitas metode deteksi *DIF*, disarankan melakukan penelitian serupa, namun dengan replikasi yang lebih besar dan persentase kandungan butir *true DIF* yang lebih lebar, misalnya, sebesar 0% – 80%. Hal ini diperlukan untuk lebih meyakinkan metode mana yang mempunyai sensitivitas paling tinggi dan metode mana yang mempunyai ketepatan paling tinggi pada populasi yang lebih luas.

g. Untuk data riil, penelitian ini hanya melibatkan satu UAN saja. Para peneliti disarankan untuk dapat melakukan penelitian serupa, tetapi dengan mengambil UAN Matematika untuk beberapa tahun sehingga pembahasan mengenai pokok-pokok bahasan yang cenderung memuat *DIF* dan sebab-sebab terjadinya *DIF* pada butir-butir soal Matematika dapat dilakukan lebih komprehensif. Peneliti juga dapat melanjutkan penelitian ini dengan mengambil beberapa mata pelajaran lain di luar Matematika dan mengguna-kan pengelompokan di luar jenis kelamin.

Daftar Pustaka

Badrun Kartowagiran (2005). *Perbandingan berbagai metode untuk mendeteksi bias butir.* Disertasi doktor, tidak diterbitkan, Universitas Gajahmada, Yogyakarta.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37*, 307-327.

Carr, M., Jessup, D. L., & Fuller, D. (1999). Gender differences in first-grade mathematics strategy uses: Parent and teacher contributions. *Journal for Research in Mathematics Education*, 30, 20-46.

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

Fennema, E. (2000). Gender and mathematics: What is known and what do I wish was known?. *Paper.* Presented at the fifth annual forum of the National Institute for Science Education. Diambil pada tanggal 23 Mei 2005 dari http://www.wccr.wisc.edu/nisc/News_Activities/_Forums.

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher, 27,* 6-11.

Field, A. (2000). *Discovering statistics using SPSS for windows: Advanced techniques for the beginner.* London: Sage Publications.

French, A. W. & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33,* 315-332.

Gierl, M. J., Khaliq, S. N. & Boughton, K. (1999). Gender differential item functioning in mathematics and science: Prevalence and policy implications. *Paper.* Presented at the annual meeting of the Canadian society for the study of education. Diambil pada tanggal 20 Januari 2003, dari http:// www.ncrel.org/sdrs.

Heri Retnawati (2003). *Keberfungsian butir diferensial pada perangkat tes seleksi masuk sekolah lanjutan tingkat pertama (SLTP) mata pelajaran matematika.* Tesis magister, tidak diterbitkan, Universitas Negeri Yogyakarta.

Holland, P. W. & Thayer, D. T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. Dalam H. Wainer & H. I. Braun (Eds), *Test Validity* (pp 129-145). Hillsdale: Lawrence Erlbaum Associates Publisher.

Hsin-Hung Li, Nandakumar, R., & Stout, W. (1995). Application of SIBTEST in dealing with issues of DIF in the context of multidimensional data. *Paper.* Presented at the annual meeting of the National Council on Measurement in Education, San Fransisco, 19 April, 1995. Diambil pada tanggal 15 Mei 2003 dari http://www.stat.uiuc.edu/stoutlab/papers.

Hulin, C. L., Drasgow, F. & Parson, C. K. (1993). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Lips, H. M. (1988). *Sex & gender: An introduction.* Mountain View, CA: Mayfield Publishing Company.

Penfield, R. D. & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*, 5-15.

Purwo Susongko (2000). *Keberfungsian butir diferensial perangkat tes Ebtanas Kimia sekolah menengah umum di Jawa Tengah.* Tesis magister, tidak diterbitkan, Universitas Negeri Yogyakarta.

Richmond-Abbot, M. (1992). *Masculine and feminine: Gender role over the life cycle (2nd ed.).* New York: McGraw Hill, Inc.

Rogers, H. J. & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied psychological measurement, 17*, 105-116.

Roussos, L. A., Schnipke, D. L. & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of educational and behavioral statistics, 24*, 293-322.

Rudas, T. & Zwick, R. (1997). Estimating the importance of differential item functioning. *Journal of educational and behavioral statistics, 22*, 31-45.

Shealy, R. T. & Stout, W. F. (1993). An item responses theory model for test bias and differential test functioning. Dalam P. W. Holland & H. Wainer (Eds), *Differential item functioning.* (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum Associates Publisher.

Stafslien, C. (2001). *Gender differences in achievement in mathematics.* Diambil pada tanggal 23 Mei 2005 dari http://www.math.wisc.edu/ ~weinberg/ MathEd.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of educational measurement. 27.* 361 – 370.

Unger, R. & Crawford, M. (1992). *Women and gender: A feminist psychology.* New York: McGraw Hill, Inc.