Jurnal Evaluasi Pendidikan
Volume 7, Nomor 2, Oktober 2016
DOI: doi.org/10.21009/JEP.072.01

SENSITIVITAS MENDETEKSI BIAS BUTIR METODE UJI BEDA TARAF SUKAR, KHI-KUADRAT LORD DAN DISTRIBUSI SAMPLING EMPIRIS

I Wayan Widana
Institut Keguruan Ilmu Pendidikan (IKIP) PGRI Bali

ABSTRACT
The objective of this research was to understand the sensitivity difference of DIF methods: (1) test of difficulty level difference, (2) Lord's chi-square, (3) empirical sampling distributions to detect item bias. This research is of descriptive exploratory nature. Data were obtained from Puspendik in a form of response data of a high school student majoring in science in DKI Jakarta Province of 4,869 people. The item bias test being tested was the mathematics national exam of the 2014/2015 academic year with package code 1101. To detect item bias, replication is done 30 times, the samples are randomly selected from both the reference group and focal group, each consisting of 1000 people. Data were analyzed by using the Bilog-MG Program, SPSS 22.0, and Microsoft Excel 2010. The hypothesis is tested by using One_Way ANOVA statistical and Tukey Post Hoc test. The results showed that: (1) the empirical sampling distributions method is more sensitive than the Lord's chi-square method to detect item bias with mean difference of 3.200 and value-sig.0.000<0.05; (2) the Lord's chi-square method was more sensitive than the test of b difference method for detecting item bias, with mean difference of 2.367 and value-sig.0.001<0.05.

Keywords
Test of b difference, Lord's chi-square, empirical sampling distribution, item bias

ABSTRAK
Penelitian ini bertujuan untuk mengetahui perbedaan sensitivitas metode *DIF*: (1) Uji Beda Taraf Sukar, (2) Khi-Kuadrat Lord, dan (3) Distribusi Sampling Empiris. Penelitian ini merupakan penelitian deskriptif eksploratif. Data diperoleh dari Puspendik, berupa data respons siswa SMA Program IPA Provinsi DKI Jakarta yang berjumlah 4.869 orang. Butir soal yang dianalisis adalah butir soal UN Matematika tahun pelajaran 2014/2015 kode paket 1101. Untuk mendeteksi bias butir dilakukan replikasi secara acak sebanyak 30 kali, pada kelompok fokal dan referensi masing-masing berjumlah 1000 orang. Data dianalisis menggunakan Program BILOG-MG, SPSS 22.0, dan Excel 2010. Pengujian hipotesis menggunakan statistik *One_Way ANOVA* dan uji *Post Hoct* Tukey. Hasil penelitian menunjukkan bahwa: (1) metode Distribusi Sampling Empiris lebih sensitif mendeteksi bias butir daripada metode Khi-Kuadrat Lord dengan perbedaan rerata 3.200 dan nilai sig. 0.000<0.05; (2) metode Khi-Kuadrat Lord lebih sensitif mendeteksi bias butir daripada metode Uji Beda Taraf Sukar dengan perbedaan rerata 2.367 dan nilai sig. 0.001<0.05.

Kata Kunci
Uji beda taraf sukar, Khi-kuadrat Lord, distribusi sampling empiris, bias butir

Alamat Korespondensi
Institut Keguruan Ilmu Pendidikan (IKIP) PGRI, Bali
Indonesia
e-mail:
iwyn_widana@yahoo.co.id

1. Pendahuluan

Peraturan Pemerintah Nomor 13 Tahun 2015 tentang Perubahan Kedua Atas Peraturan Pemerintah Republik Indonesia Nomor 19 Tahun 2005 tentang Standar Nasional Pendidikan (SNP), pasal (1) menyatakan bahwa evaluasi pendidikan adalah kegiatan pengendalian, penjaminan, dan penetapan mutu pendidikan terhadap berbagai komponen pendidikan pada setiap jalur, jenjang, dan jenis pendidikan sebagai bentuk pertanggungjawaban penyelenggaraan pendidikan. Oleh karena itu, perlu ditetapkan standar penilaian pendidikan yang terkait dengan kriteria mengenai mekanisme, prosedur, dan instrumen penilaian hasil belajar peserta didik.

Lebih lanjut Peraturan Menteri Pendidikan dan Kebudayaan Nomor 53 Tahun 2015 tentang Penilaian Hasil Belajar oleh Pendidik pada Pendidikan Dasar dan Pendidikan Menengah, pasal

(4), menyatakan bahwa untuk menjamin agar penilaian dapat memberikan informasi yang benar, maka penilaian hendaknya dilakukan menggunakan prinsip-prinsip penilaian antara lain: (1) sahih, berarti penilaian didasarkan pada data yang mencerminkan kemampuan yang diukur; (2) objektif, berarti penilaian didasarkan pada prosedur dan kriteria yang jelas, tidak dipengaruhi subjektivitas penilai; (3) adil, berarti penilaian tidak menguntungkan atau merugikan peserta didik karena berkebutuhan khusus serta perbedaan latar belakang agama, suku, budaya, adat istiadat, status sosial ekonomi, dan gender; (4) terpadu, berarti penilaian oleh pendidik merupakan salah satu komponen yang tak terpisahkan dari kegiatan pembelajaran; (5) terbuka, berarti prosedur penilaian, kriteria penilaian, dan dasar pengambilan keputusan dapat diketahui oleh pihak yang berkepentingan; (6) menyeluruh dan berkesinambungan, berarti penilaian oleh pendidik mencakup semua aspek kompetensi dan dengan menggunakan berbagai teknik penilaian yang sesuai dengan kompetensi yang harus dikuasai peserta didik; (7) sistematis, berarti penilaian dilakukan secara berencana dan bertahap dengan mengikuti langkah-langkah baku; (8) beracuan kriteria, berarti penilaian didasarkan pada ukuran pencapaian kompetensi yang ditetapkan; dan (9) akuntabel, berarti penilaian dapat dipertanggungjawabkan baik dari segi teknik, prosedur, maupun hasilnya.

Sesuai dengan prinsip-prinsip penilaian di atas, Pusat Penilaian Pendidikan (2013) menyebutkan terdapat dua permasalahan pokok yang harus dipertimbangkan untuk mendapatkan hasil pengukuran yang baik, yaitu: alat ukur atau tes yang digunakan dapat menghasilkan skor yang reliabel, dan alat ukur tersebut harus valid, artinya dapat mengukur dengan tepat sesuai dengan target objek (kemampuan peserta tes) yang akan diukur. Apabila alat ukur telah memenuhi kedua kriteria tersebut, akan diperoleh hasil pengukuran ulang yang sesuai dengan tujuan pengukuran tanpa terpengaruh oleh faktor-faktor lainnya. Hasil pengukuran diharapkan dapat memberikan informasi yang sesuai dengan karakteristik peserta tes tanpa merugikan atau menguntungkan individu atau kelompok-kelompok tertentu akibat ketidakadilan alat ukur tersebut.

Di dalam kenyataan, ada kalanya skor hasil tes itu tidak memberikan informasi yang benar terhadap karakteristik peserta tes. Mungkin saja informasi tersebut tidak menjangkau sampai ke besaran atau dimensi yang hendak diukur oleh tes tersebut. Mungkin pula hasil tes tersebut tercampur dengan besaran atau dimensi lain yang tidak dimaksudkan untuk diukur oleh tes tersebut sehingga hasil tes menjadi rancu. Akibatnya skor yang diperoleh tidak benar atau terjadi ketimpangan skor. Ketimpangan skor tidak memberikan informasi yang benar tentang hal-hal yang dimaksud untuk diukur oleh uji tes itu (Naga, 1992). Dalam penilaian, ketimpangan skor hendaknya dapat diminimalkan agar pengukuran dapat memberikan gambaran yang utuh tentang karakteristik objek yang diukur.

Jika pada suatu tes memuat butir-butir yang memihak kelompok tertentu, maka tes tersebut dikatakan memuat bias butir atau mengandung keberfungsian butir diferensial (*Differential Item Functioning, DIF*). Bias butir hendaknya diminimalkan atau dihilangkan sama sekali. Terutama pada tes yang digunakan secara luas dan hasilnya sangat menentukan masa depan seseorang seperti Ujian Sekolah, UN, SBMPTN, dan berbagai bentuk tes seleksi lainnya. Tes yang digunakan harus bebas dari bias butir yang dapat merugikan individu atau sekelompok individu tertentu. Sebelum tes beserta butir-butirnya digunakan, seharusnya diujicobakan terlebih dahulu sehingga apabila terdapat butir-butir yang mengandung bias dapat terdeteksi sejak awal.

Menurut Naga (1992) untuk mendeteksi adanya bias butir dalam seperangkat tes, terdapat sejumlah metode yang dapat digunakan. Dalam teori klasik dikenal beberapa metode antara lain: (1) kelompok tunggal (*single group validity*), (2) korelasi diferensial (*differential validity*), (3) model regresi atau model *Clearly*, (4) prosedur diskriminasi butir (*item discrimination procedure*), (5) metode *plot delta* (*delta plot method*), (6) pendekatan khi-kuadrat *Scheuneman (Scheuneman chi-squared approach)*, (7) pendekatan khi-kuadrat Camilli *(Camilli chi-squared approach).* Berbeda dengan pendekatan klasik, pendekatan modern (*IRT*) memiliki sifat invarian. Skor peserta invarian terhadap ubahan pada butir tes, serta skor butir invarian terhadap ubahan peserta tes. Secara garis besar ada dua cara yang paling banyak digunakan

untuk pendeteksian *DIF,* yaitu: (1) pencocokan parameter ciri butir di antara sub populasi 1 dan sub populasi 2 sambil memeriksa kecocokan model; (2) penghitungan luas wilayah di antara lengkungan karakteristik butir yang dibentuk oleh subpopulasi 1 dan subpopulasi 2. Butir tidak bias jika luas itu sama dengan nol atau sangat kecil.

Sementara itu, Camilli dan Shepard (1994) mengemukakan bahwa terdapat 5 metode untuk mendeteksi bias butir dengan pendekatan teori responsi butir, yaitu: (1) *test of b difference* (uji beda taraf sukar), (2) *item drift methods*, (3) *Lord's chi-square*, (4) *empirical sampling distributions* (distribusi sampling empiris), dan (5) *model comparison measures*.

Di antara metode *DIF* yang menggunakan pendekatan teori responsi butir yang telah dipaparkan di atas, maka yang digunakan untuk mendeteksi bias butir dalam penelitian ini adalah metode Uji Beda Taraf Sukar, Khi-Kuadrat Lord, dan Distribusi Sampling Empiris. Pemilihan tersebut didasarkan pada rekomendasi oleh Camilli dan Shepard, yang menyatakan bahwa metode-metode *DIF* yang berdasarkan teori responsi butir lebih akurat dibandingkan dengan metode *DIF* yang didasarkan pada teori tes klasik. Di samping itu ketiga metode *DIF* Uji Beda Taraf Sukar, Khi-Kuadrat Lord, dan Distribusi Sampling Empiris penyamaan skala dan proses mendeteksi *DIF* dilakukan secara terpisah (*sparate*). Dengan demikian, penelitian ini membandingkan sensitivitas metode *DIF* menggunakan pendekatan teori responsi butir.

Teori tes modern (*IRT*) dikembangkan untuk mengatasi berbagai keterbatasan dalam teori tes klasik. Hasil yang diperoleh melalui pendekatan teori klasik dirasakan kurang akurat. Naga (2013) mengemukakan bahwa teori responsi butir berusaha meningkatkan akurasi pengukuran melalui pemisahan parameter butir dari kemampuan responden. Berapapun kemampuan responden, nilai parameter butir tidak berubah. Karakteristik butir ditentukan oleh responsi para responden (baik kemampuan tinggi maupun kemampuan rendah) sehingga dikenal sebagai teori responsi butir (*Item Response Theory*).

Dalam *IRT* model distribusi yang digunakan adalah distribusi logistik. Terdapat 3 (tiga) model logistik dalam teori responsi butir, yaitu: model logistik satu parameter (L1P), model logistik dua parameter (L2P), dan model logistik tiga parameter (L3P). Perbedaan di antara ketiga model tersebut terletak pada banyaknya parameter yang digunakan untuk menggambarkan karakteristik butir pada model yang digunakan. Parameter-parameter yang digunakan adalah taraf sukar butir b, daya pembeda a, dan kebetulan menjawab betul c. Secara matematis model logistik dalam *IRT* dapat dirumuskan sebagai berikut.

Model L1P:

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}$$

Model L2P:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

Model L3P:

$$P_i(\theta) = c_i + (1c_i)\,\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

Hambleton dan Swaminathan (1985) menyatakan bahwa dalam *IRT* selain terpenuhinya asumsi unidimensi, invariansi parameter, dan independensi lokal, hal penting yang perlu diperhatikan adalah pencocokan model. Pencocokan model dilakukan dengan tujuan untuk memeriksa kecocokan antara model karakteristik yang dipilih dengan data dari lapangan. Dalam pemeriksaan ini kemungkinan ada butir yang cocok dan ada butir yang tidak cocok. Apabila model telah ditetapkan maka semua perhitungan akan berkisar di sekitar model itu. Sehingga bila data lapangan cocok dengan model karakteristik butir yang dipilih, maka estimasi terhadap parameter dapat dilakukan dengan akurat. Di samping itu, pada butir-butir yang cocok dengan model berarti syarat-syarat unidimensi, independensi lokal, dan invarian parameter telah terpenuhi untuk masing-masing butir. Dengan demikian, hanya butir-butir yang cocok dengan model yang dapat dianalisis menggunakan *IRT*.

Fungsi informasi adalah informasi tentang hubungan di antara parameter kemampuan responden dengan parameter butir. Dalam teori responsi butir, nilai fungsi informasi berbanding terbalik dengan ketidakpastian. Hal ini berarti bahwa makin kecil ketidakpastian maka makin tinggi nlai fungsi informasinya, dan sebaliknya. Fungsi informasi dibedakan menjadi dua jenis, yaitu: 1) fungsi informasi butir, merupakan

kemampuan dan kontribusi suatu butir untuk mengungkapkan kemampuan laten (*latent traits*) yang diukur menggunakan butir soal tersebut; dan 2) fungsi informasi tes, merupakan jumlah dari fungsi informasi butir-butir soal yang ada pada tes tersebut. Akibatnya, nilai fungsi informasi tes akan tinggi bila butir-butir soal penyusun tes memiliki nilai fungsi informasi yang tinggi pula. Makin tinggi nilai fungsi informasi tes, maka makin baik mutu tes untuk mengungkapkan kemampuan laten responden (Hambleton, Swaminathan, dan Rogers, 1991).

Sebelum dilakukan pendeteksian bias butir menggunakan metode *DIF*, parameter butir harus disamakan atau paramater butir skala kelompok fokal dan kelompok referensi harus berada dalam skala yang sama (Camilli & Shepard, 1994; Won-Chan Lee & Jae-Chun Ban, 2007; Allan S. Cohen & Seok-Ho Kim, 1998). Dalam penelitian ini metode penyetaraan parameter yang digunakan adalah metode rerata dan sigma, sebagaimana direkomendasikan oleh Wardani Rahayu (2007). Metode rerata dan sigma menggunakan rerata dan simpangan baku dari estimasi taraf sukar butir dari kedua kelompok peserta tes. Metode ini sangat cocok digunakan untuk data model dikotomus.

Terdapat perbedaan pengertian antara *Differential Item Functioning* (*DIF)* dan bias butir. Namun banyak penulis menganggap kedua istilah itu memiliki pengertian yang sama dan dapat dipertukarkan (Camilli & Shepard, 1994; Likun Hou, Jimmy de la Torre, dan Ratna Nandakumar, 2014). *DIF* diartikan sebagai pengukuran taraf sukar relatif butir yang menyimpang secara berlebihan pada kelompok berbeda dan responden yang memiliki kemampuan sama. Sedangkan bias butir diartikan sebagai ketidakvalidan suatu butir, sehingga butir tersebut tidak memberikan peluang yang sama menjawab benar pada dua kelompok yang berbeda pada peserta dengan kemampuan yang sama (Osterlind, 1983; Mazor 1995). Idealnya statistik *DIF* itu digunakan untuk mencari butir tes yang berfungsi secara berbeda untuk kelompok yang berbeda kemudian setelah dilakukan analisis logis mengapa butir-butir itu secara relatif dirasakan lebih sulit, sekelompok butir bias kemudian dapat diidentifikasikan sebagai bias dan tentu saja harus dikeluarkan dari tes. Namun, ketika tidak menganalisis penyebab timbulnya bias maka kedua istilah itu dapat digunakan secara bergantian. Dengan demikian *DIF* merupakan metode statistika yang digunakan untuk mendeteksi adanya bias butir.

Butir-butir yang dikategorikan bias dapat menguntungkan salah satu kelompok untuk semua interval kemampuan. Pada konsidi lainnya, bias butir juga dapat menguntungkan peserta tes pada kelompok tertentu hanya pada level kemampuan tertentu saja. Terdapat dua jenis bias butir, yaitu: bias butir *uniform* (konsisten) dan bias butir *non-uniform* (tidak konsisten). Finh & French (2007) mengemukakan bahwa bias butir *uniform* muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya terjadi pada setiap level kemampuan, sedangkan bias butir *non-uniform* muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya tidak terjadi pada setiap level kemampuan. Dilihat dari kurva karakteristik butir, pada bias butir *uniform* (konsisten) grafik kelompok yang satu selalu berada di bawah kelompok yang lainnya. Sedangkan pada bias butir *non-uniform,* grafik kelompok yang satu tidak selalu berada di bawah kelompok yang lainnya dengan kata lain kedua kurva saling berpotongan.

Yanmei Li, Allan S. Cohen, dan Robert A. Ibarra (2004) menyelidiki karakteristik butir soal matematika ditinjau dari perbedaan gender. Bias butir dapat terjadi karena adanya perbedaan gender dan etnik. Perbedaan gender menyebabkan adanya perbedaan strategi dalam penyelesaian soal-soal matematika. Problem matematika dapat diklasifikasikan menjadi problem konvensional dan nonkonvensional. Problem konvensional seperti problem rutin (*textbook*) yang dapat diselesaikan dengan algoritma matematika biasa, sedangkan problem nonkonvensional umumnya memerlukan cara penyelesaian yang tidak umum dilakukan, memerlukan wawasan, serta cara berpikir kritis dan kreatif. Hasil penelitiannya menyatakan bahwa perempuan umumnya lebih unggul dibandingkan dengan laki-laki dalam menyelesaikan masalah konvensional yang mengikuti pola keteraturan (algoritma). Sebaliknya, laki-laki lebih unggul daripada perempuan untuk menyelesaikan masalah nonkonvensional yang menuntut kemampuan berpikir tingkat tinggi (*higher order*

thinking skills). Perempuan juga lebih baik dalam hal kecepatan mengakses informasi dan menyelesaikan soal-soal yang memerlukan ingatan, sedangkan laki-laki lebih unggul dalam menyelesaikan problem matematika yang menuntut inovasi dan kreativitas tinggi.

Dari uraian di atas dapat disimpulkan bahwa perbedaan jenis kelamin dan gender dapat mempengaruhi kemampuan seseorang untuk memecahkan suatu masalah. Peserta didik laki-laki dan perempuan memiliki sejumlah perbedaan yang disebabkan oleh adanya perbedaan perlakuan, kebiasaan, perasaan, dan kebudayaan antara etnis satu dengan yang lainnya. Di samping itu, pada umumnya peserta didik laki-laki dilihat dari postur tubuhnya lebih kuat bila dibandingkan dengan peserta didik perempuan. Oleh karena itu, apabila peserta didik laki-laki dan perempuan diberikan menjawab sejumlah butir soal yang sama, dapat terjadi perbedaan peluang menjawab benar antar kedua kelompok tersebut. Kondisi tersebut dapat menyebabkan terjadi bias butir karena perbedaan jenis kelamin.

Metode *DIF* yang diperbandingkan dalam penelitian ini adalah metode *DIF* yang berdasarkan teori responsi butir. Adapun metode *DIF* yang dibandingkan adalah: 1) Uji Beda Taraf Sukar, 2) Khi-kuadrat Lord, dan 3) Distribusi Sampling Empiris. Teknik analisis masing-masing metode DIF yang diperbandingkan dipaparkan sebagai berikut.

Untuk mendeteksi bias butir pada metode Uji Beda Taraf Sukar, terlebih dahulu harus ditentukan *standard error* dari $\hat{b}_F$ dan $\hat{b}_R$. Dengan $\hat{b}_F$ merupakan estimasi parameter taraf sukar kelompok fokal dan $\hat{b}_R$ merupakan estimasi parameter taraf sukar kelompok referensi. Bila S_F dan S_R berturut-turut merupakan *standard error* dari $\hat{b}_F$ dan $\hat{b}_R$, maka *standard error* perbedaan taraf sukar

$$\Delta\hat{b} = \hat{b}_F - \hat{b}_R$$

dilambangkan dengan:

$$S_{\Delta\hat{b}} = \sqrt{S_F^2 + S_R^2}$$

Statistik yang digunakan untuk menguji signifikansi uji beda taraf sukar butir adalah $= \frac{\Delta\hat{b}}{S_{\Delta\hat{b}}}$. Jika nilai d > z = 1,96 atau d < z = -1,96 pada taraf signifikansi $\alpha = 0{,}05$,maka butir ke-i dinyatakan memuat bias butir. Pada metode Khi-Kuadrat Lord, bias butir dideteksi menggunakan rumus: $\chi_i^2 = \nu_i' \sum_i^{-1} \nu_i$. Jika χ^2_{hitung} > 7,81472 dengan dk=3, $\alpha = 0{,}05$, maka butir ke-i dinyatakan memuat bias butir. Sedangkan pada metode DSE, terlebih dahulu dihitung indeks bias butir antar kelompok R-F (δ) menggunakan rumus: $SPD - \theta = \frac{\sum_{j=1}^{np} \Delta P(\theta_j)}{n_p}$ untuk bias butir *uniform* dan $UPD - \theta = \sqrt{\frac{\sum_{j=1}^{np} (\Delta P(\theta_j))^2}{n_p}}$ untuk bias butir *non-uniform*, dengan: $\Delta P = P_R(\theta_j) - P_F(\theta_j)$. Selanjutnya dihitung indeks bias butir subkelompok R1-R2 (δ_1), dan F1-F2 (δ_2). Jika nilai δ >nilai maksimum (δ_1; δ_2), maka butir ke-i dinyatakan memuat bias butir.

Tujuan penelitian ini untuk mengetahui: (1) perbandingan sensitivitas metode *DIF* untuk mendeteksi bias butir pada metode Distribusi Sampling Empiris dan Khi-Kuadrat Lord; dan (2) perbandingan sensitivitas metode *DIF* untuk mendeteksi bias butir pada metode Khi-Kuadrat Lord dan Uji Beda Taraf Sukar.

2. Metode Penelitian

Metode yang digunakan di dalam penelitian ini adalah metode eksperimen dengan populasi terdiri atas dua jenis, yaitu: populasi peserta tes dan populasi butir. Populasi peserta tes adalah peserta didik SMA Negeri dan Swasta di Provinsi DKI Jakarta peserta UN tahun pelajaran 2014/2015, yang dipilah menjadi populasi peserta didik laki-laki dan peserta didik perempuan. Populasi peserta tes berjumlah 4.869 orang peserta didik, yang terdiri dari 2.114 orang peserta didik laki-laki dan 2.755 peserta didik perempuan. Sampel peserta tes diambil secara acak dari masing-masing populasi, yaitu: kelompok perempuan sebagai kelompok fokal dan kelompok laki-laki sebagai kelompok referensi. Jumlah sampel responden yang akan digunakan dalam penelitian ini adalah 2.000 orang. Masing-masing 1.000 orang untuk kelompok referensi dan 1.000 orang untuk kelompok fokal. Hal ini sesuai dengan pendapat Naga (1992) yang menyatakan bahwa ukuran sampel minimal pada teori responsi butir sangat tergantung pada

model yang digunakan. Untuk model L3P ukuran sampelnya lebih besar daripada model L2P, ada ukuran sampel yang kurang dari 500, ada yang kurang dari 1.000, ada pula yang kurang dari 2.000. Makin besar ukuran sampel, makin baik hasil estimasi model respons karena berkaitan dengan uji kecocokan model.

Sedangkan populasi butir dalam penelitian ini adalah butir-butir soal Matematika Program IPA, yang terdapat dalam perangkat tes UN Provinsi DKI Jakarta tahun pelajaran 2014/2015 dalam bentuk paket-paket soal. Sampel butir adalah butir-butir soal UN mata pelajaran Matematika Program IPA sebanyak 40 butir, pada paket dengan kode 1101.

Data dalam penelitian ini adalah data hasil UN mata pelajaran Matematika Program IPA Provinsi DKI Jakarta tahun pelajaran 2014/2015. Data tersebut diperoleh dari Puspendik Balitbang Kemdikbud. Data tersebut berupa respons peserta didik dalam format *Excel*. Mengingat kebijakan yang dilakukan oleh Puspendik tidak mengeluarkan kunci jawaban, maka kunci jawaban dibuat oleh peneliti bersama guru-guru Matematika SMA senior dan dosen Matematika dari Perguruan Tinggi. Respons peserta didik yang awalnya dalam bentuk huruf dikonversi dalam bentuk angka, yaitu: untuk jawaban benar diberi angka 1 dan jawaban salah diberi angka 0, sehingga dapat dilakukan penskoran.

Untuk menguji perbedaan sensitivitas masing-masing metode *DIF* dilakukan melalui dua tahap analisis, sebagai berikut: (1) tahap pertama, menguji perbedaan rerata jumlah butir yang mengandung bias butir pada ketiga metode *DIF* menggunakan anava satu jalur; (2) tahap kedua, apabila secara statistika terbukti terdapat perbedaan rerata jumlah butir yang mengandung bias butir pada ketiga metode *DIF,* maka dilanjutkan dengan uji *Post Hoc*, menggunakan uji Tukey (karena jumlah sampel yang diuji sama). Analisis data menggunakan Program SPSS 22.0.
Kriteria pengujian:

(1) Pada ***output*** SPSS 22.0, uji hipotesis dibaca pada tabel ***ANOVA***. Apabila nilai ***F*** dengan nilai ***sig < 0,05*** (signifikan), maka tolak H_0 dan terima H_1 yang berarti bahwa terdapat perbedaan yang signifikan rerata jumlah bias butir pada ketiga metode *DIF*. Sebaliknya, jika nilai ***F*** dengan nilai ***sig > 0,05*** (tidak signifikan), maka tolak H_0 dan terima H_1 yang berarti bahwa tidak terdapat perbedaan yang signifikan rerata jumlah bias butir pada ketiga metode *DIF*.

(2) Apabila terbukti terdapat perbedaan yang signifikan rerata jumlah bias butir pada ketiga metode *DIF*, maka dilanjutkan uji *Post Hoc* menggunakan Uji Tukey. Uji ***Post Hoc*** dapat dibaca pada tabel ***Multiple Comparisons, output*** SPSS 22.0. Perbedaan rerata masing-masing metode *DIF* dapat dibaca pada kolom *Mean Difference (I-J)*. Jika nilai *Mean Difference (I-J)* dengan nilai sig.p < 0,05 (signifikan), berarti bahwa perbedaan rerata banyaknya bias butir pada metode *DIF* ke-i dan metode *DIF* ke-j signifikan. Dengan demikian metode *DIF* ke-i dikatakan lebih sensitif daripada metode *DIF* ke-j. Sebaliknya, jika nilai *Mean Difference (I-J)* dengan nilai sig.p > 0,05 (tidak signifikan), berarti bahwa perbedaan rerata banyaknya bias butir pada metode *DIF* ke-i dan metode *DIF* ke-j tidak signifikan.

3. Hasil Penelitian dan Pembahasan

Untuk melakukan analisis bias butir menggunakan tiga metode *DIF,* yaitu: (a) metode Uji Beda Taraf Sukar (UBTS), (b) Khi-Kuadrat Lord, dan (c) Distribusi Sampling Empiris (DSE), analisis dilakukan dengan 2 tahap. Tahap pertama analisis parameter butir, matriks varians-kovarians, dan parameter kemampuan (θ) untuk masing-masing kelompok referensi dan kelompok fokal serta subkelompoknya dianalisis menggunakan Program BILOG-MG 3.07. Tahap kedua, perhitungan analisis bias butir untuk ketiga metode *DIF* yang diperbandingkan menggunakan Program Excel 2010. Untuk mendeteksi adanya bias butir pada masing-masing metode *DIF* dilakukan sebanyak 30 replikasi. Banyaknya bias butir pada masing-masing metode DIF pada 30 kali replikasi.

Uji hipotesis dilakukan dengan menggunakan Program SPSS 22.0. Pengujian hipotesis dilakukan melalui 2 tahap sebagai berikut.

1. Tahap pertama, menguji perbedaan rerata jumlah soal bias butir pada ketiga model *DIF* menggunakan ***One-Way ANOVA***. Hipotesis yang diuji adalah sebagai berikut.

   $H_0 \quad : \quad \mu_A = \mu_B = \mu_C$

H_1 : Bukan H_0

Keterangan:

μ_A : Rerata banyak bias butir pada metode Distribusi Sampling Empiris

μ_B : Rerata banyak bias butir pada metode Khi-Kuadrat Lord

μ_C : Rerata banyak bias butir pada metode Uji Beda Taraf Sukar

Tabel 1. Uji Perbedaan Rerata ANOVA

	Sum of squere	df	Mean squere	f	Sig.
Between Groups	468.289	2	234.144	38.452	.000
Within Groups	529.767	87	6.089		
Total	998.056	89			

Data pada Tabel 1, menunjukkan bahwa nilai F sebesar 38,452 dengan signifikansi sig. = 0,000 < 0,05 (signifikan). Kesimpulan: tolak H_0 dan terima H_1. Hal itu berarti bahwa terdapat perbedaan yang signifikan rerata jumlah butir bias yang dideteksi menggunakan metode UBTS, Khi-Kuadrat Lord, dan Distribusi Sampling Empiris. Selanjutnya, untuk mengetahui perbandingan sensitivitas metode *DIF* dilanjutkan dengan melakukan uji *Post Hoc*.

2. Tahap kedua, uji *Post Hoc* juga dilakukan menggunakan Program SPSS 22.0. Hipotesis yang diuji adalah sebagai berikut.

(1) $H_0 : \mu_A - \mu_B = 0$

$H_1 : \mu_A - \mu_B > 0$

(2) $H_0 : \mu_B - \mu_C = 0$

$H_1 : \mu_B - \mu_C > 0$

Tabel 2. Hasil Uji *Post Hoc*

Multiple Comparisons

Dependent Variable: BIAS_BUTIR

Tukey HSD

(I) METODE	(J) METODE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
UBTS	LORD	-2.367*	.637	.001	-3.89	-.85
	DSE	-5.567*	.637	.000	-7.09	-4.05
LORD	UBTS	2.367*	.637	.001	.85	3.89
	DSE	-3.200*	.637	.000	-4.72	-1.68
DSE	UBTS	5.567*	.637	.000	4.05	7.09
	LORD	3.200*	.637	.000	1.68	4.72

*. The mean difference is significant at the 0.05 level.

Data hasil uji ***Post Hoc*** pada Tabel 2, dapat dijelaskan sebagai berikut:

a. Metode DSE dan metode Khi-Kuadrat Lord memiliki perbedaan rerata (*Mean Difference*) sebesar 3,200 dengan nilai sig.0,000<0,05 (signifikan), kesimpulan: tolak H_0 dan terima H_1. Hal itu berati bahwa metode DSE lebih sensitif mendeteksi bias butir dibandingkan dengan metode Khi-Kuadrat Lord. Hipotesis pertama terbukti.

b. Metode Khi-Kuadrat Lord dan metode UBTS memiliki perbedaan rerata (*Mean Difference*) sebesar 2,367 dengan nilai sig.0,001<0,05 (signifikan), kesimpulan: tolak H_0 dan terima H_1. Hal itu berati bahwa metode Khi-Kuadrat Lord lebih sensitif mendeteksi bias butir dibandingkan dengan metode UBTS. Hipotesis kedua terbukti.

Sebagian besar butir-butir yang dinyatakan bias oleh metode UBTS juga dinyatakan bias secara keseluruhan (sempurna) oleh metode Khi-Kuadrat Lord dan DSE. Namun terdapat juga sebagian kecil butir-butir yang dinyatakan bias oleh metode *DIF* yang satu, tetapi tidak terdeteksi bias oleh metode *DIF* lainnya. Misalnya terdapat

butir-butir yang dinyatakan bias oleh metode UBTS, tidak terdeteksi oleh metode Khi-Kuadrat Lord, tetapi terdeteksi oleh metode DSE. Atau terdapat pula butir-butir yang dinyatakan bias oleh metode UBTS dan Khi-Kuadrat Lord, tetapi tidak terdeteksi oleh metode DSE. Perbedaan tersebut diakibatkan oleh adanya perbedaan teknik dalam perhitungan analisis bias butir pada masing-masing metode *DIF*.

Metode UBTS kurang peka untuk mendeteksi bias butir bila dibandingkan dengan metode Khi-Kuadrat Lord, disebabkan metode UBTS yang mengasumsikan parameter daya pembeda sama dan faktor tebakan diasumsikan nol. Sedangkan metode Khi-Kuadrat Lord memperhatikan semua parameter butir dan matriks varians-kovarians. Perbedaan nilai estimasi parameter a mengakibatkan perbedaan kecuraman kurva karakteristik butir. Sedangkan perbedaan nilai estimasi parameter b, mengakibatkan pergeseran kurva karakteristik butir ke kiri atau ke kanan. Makin besar nilai b (butir tersebut makin sulit), maka posisi kurva karakteristik butir makin ke kanan. Sedangkan perbedaan nilai estimasi parameter c, mengakibatkan pergeseran kurva karakteristik butir ke atas atau ke bawah. Makin besar nilai c (tebakan makin tinggi), maka posisi kurva karakteristik butir makin ke atas.

Pada metode DSE, untuk mendeteksi bias butir menggunakan selisih probabilitas menjawab benar dari sebuah butir soal yang diberikan kepada peserta dengan kemampuan yang sama pada kelompok yang berbeda. Secara teoretis, keunggulan metode DSE dibandingkan dengan metode Khi-Kuadrat Lord, sebagaimana diungkapkan oleh Shepard, dkk (1994) bahwa untuk mendeteksi bias butir dilakukan melalui replikasi analisis, menggunakan sampel berpasangan yang ekuivalen random. Pada metode Khi-Kuadrat Lord, signifikansi terhadap bias butir diperhitungkan melalui perbedaan estimasi parameter butir dan matriks varians kovarians kelompok referensi dan fokal, tanpa adanya replikasi analisis seperti dalam metode DSE. Perhitungan tersebut lebih cermat bila dibandingkan dengan metode Khi-Kuadrat Lord, karena dalam perhitungannya hanya menggunakan selisih estimasi parameter butir yang dinyatakan dalam bentuk perkalian matriks. Dalam perhitungan metode Khi-Kuadrat Lord tidak mempertimbangkan parameter θ (kemampuan peserta), dan tidak melakukan replikasi analisis. Padahal sesuai dengan definisi bias butir, yaitu: butir yang memberikan peluang tidak sama untuk menjawab benar terhadap butir tertentu pada kemampuan yang sama hanya karena berasal dari kelompok yang berbeda. Sehingga metode Khi-Kuadrat Lord kurang cermat mendeteksi bias butir karena mengabaikan parameter θ, dan tidak melakukan replikasi analisis.

4. Kesimpulan

Berdasarkan hasil penelitian ditemukan Metode Distribusi Sampling Empiris (DSE) lebih sensitif daripada metode Khi-Kuadrat Lord untuk mendeteksi bias butir. Metode Khi-Kuadrat Lord lebih sensitif daripada metode Uji Beda Taraf Sukar (UBTS) untuk mendeteksi bias butir.

5. Daftar Pustaka

Allan S. Cohen & Seok-Ho Kim. (1998). An Investigation of Linking Method Under the Graded Response Model. *Journal Aplied Psychological Measurement* 22(2).

Camilli, G. dan Shepard, L. A. (1994). *Methods for Identifying Biased Test Item.* California: Sage Publications Inc.

Finch, W. Holmes dan Brian F. French. (2007). Detection of Crossing Differential Item Functioning; A Comparison of Four Methods.*Journal Education and Psychological Measurement*, 67(4): 565-567.

Hambleton, R. K., dan Swaminathan H. (1985). *Item Response Theory*. Boston: Kluwer Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. CA: Sage Publication Inc.

Likun Hou, Jimmy de la Torre, dan Ratna Nandakumar. (2014). Differential Item Functioning Assessment in Cognitive Diagnostic Modeling; Application of the Wald Test to Investigate DIF in the DINA

Model. *Journal of Educational Measurement Spring*, 51(1): 98–125.

Mazor, K. M. (1995). Using Logistic Regression and Mantel Haenszel With Multiple Ability Estimates to Detect Differential Item Funcioning. *Journal of Education Measurement. 32(2)*.

Naga, Dali S. (1992). *Pengantar Teori Sekor pada Pengukuran Pendidikan.* Jakarta: Besbats.

Naga, Dali S. (2013). *Teori Sekor pada Pengukuran Mental*. Jakarta: PT. Nagarani Citrayasa.

Osterlind, S. J. (1983). *Test Item Bias*. Baverly Hill: Sage Publication Inc.

Peraturan Pemerintah Republik Indonesia Nomor 13 Tahun 2015 tentang Perubahan Kedua Atas Peraturan Pemerintah Republik Indonesia Nomor 19 Tahun 2005 tentang Standar Nasional Pendidikan (SNP), pasal 1.

Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 53 Tahun 2015 tentang Penilaian Hasil Belajar oleh Pendidik pada Pendidikan Dasar dan Pendidikan Menengah, *pasal 4*.

Pusat Penilaian Pendidikan. (2013). *Panduan Pengembangan dan Pemberdayaan Bank Soal* di Daerah. Jakarta: Puspendik.

Rahayu, Wardani. (2007). Pengaruh Metode Linking terhadap Banyak Butir False Positive pada Pendeteksian *DIF* Berdasarkan Teori Responsi Butir. *Disertasi.* Jakarta: UNJ.

Won-Chan Lee & Jae-Chun Ban. (2007). *Comparison of Three IRT Linking Procedures in the Random Groups Equating Design*. Iowa: CASMA.

Yanmei Li, Allan S. Cohen, dan Robert A. Ibarra. (2004). Characteristics of Mathematics Items Associated with Gender DIF. *International Journal of Testing,* 4(2)