Jurnal Evaluasi Pendidikan Vol. 4, No. 2, Oktober 2013, 198-208

DETECTION OF GRAIN INSTRUMEN SCOORING WITH CORRECT SCORE AND PUNISHMENT SCORE

Wardani Rahayu
Jurusan Matematika FMIPA Universitas Negeri Jakarta
Jl. Rawamangun Muka, Jakarta
wardani9164@yahoo.com

Yuliatri Sastra Wijaya
Fakultas Teknik, Universitas Negeri Jakarta
Jl. Rawamangun Muka, Jakarta
yuliatri_s@yahoo.com

Abstract

The aim of this study was to determine the number of Differential Item Functioning (DIF) in correct score and punishment score scoring model based on sample size. DIF detection using Lord Chi Square and score equating using sigma and mean method. The amount of DIF scors for each scoring model was obtained from 30 repetition. Data analisis was using two factor Anava. The result of this study was the bigger amount of sample size then the less of DIF score detected using Lord Chi Square in correct score scoring model and the less amount of sample size then the more DIF score detected using Lord Chi Square methode in correct score scoring model

Keywords**:** *correct score*, *punishment score, DIF*, *sample size, Lord Chi Square methode*

DETEKSI DIF PADA BUTIR INSTRUMEN DENGAN PENSKORAN *CORRECT SCORE* DAN *PUNISHMENT SCORE*

Wardani Rahayu,
Jurusan Matematika FMIPA Universitas Negeri Jakarta
Jl. Rawamangun Muka, Jakarta
wardani9164@yahoo.com

Yuliatri Sastra Wijaya
Fakultas Teknik, Universitas Negeri Jakarta
Jl. Rawamangun Muka, Jakarta
yuliatri_s@yahoo.com

Abstrak

Tujuan penelitian ini adalah untuk mengetahui banyaknya butir *Differential Item Fungtioning* (*DIF*) pada model penskoran *correct score* dan *punishment score* berdasarkan ukuran sampel. Deteksi *DIF* dengan menggunakan metode *Lord Chi Square* dan penyetaraan skor dengan metode rerata dan sigma. Banyaknya butir *DIF* pada setiap model penskoran diperoleh dari 30 pengulangan. Analisis data dengan Anava dua faktor. Hasil penelitian ini adalah semakin besar ukuran sampel maka semakin sedikit butir DIF yang terdeteksi dengan metode *Lord Chi Square* pada model penskoran *correct score* dan semakin kecil ukuran sampel maka semakin banyak butir DIF yang terdeteksi dengan metode *Lord Chi Square* pada model penskoran *correct score.*

Kata kunci: *correct score*, *punishment score, DIF*, ukuran sampel, metode *Lord Chi Square*

PENDAHULUAN

Perangkat tes hasil belajar yang berbentuk pilihan ganda diantaranya tes Ujian Nasional (UN), Seleksi Perguruan Tinggi, *the Trend in International Mathematics and Science Studies* (TIMMS), *Indonesian National Asessment Program* (INAP). Bentuk tes ini dapat mengukur lebih efektif dari pada tes bentuk uraian pada skala yang besar. Bentuk tes pilihan ganda dapat mengukur cakupan materi yang cukup luas meliputi ranah kognitif pengetahuan, pemahaman, penerapan, analisis, sintesis dan evaluasi. Pada perangkat tes pilihan ganda diperlukan pengecoh yang berfungsi dengan baik sehingga mengurangi tebakan yang dilakukan oleh peserta tes. Oleh karena itu ada beberapa tes seleksi menerapkan model penskoran *punishment score* untuk mengurangi tebakan pada tes pilihan ganda seperti tes Seleksi Masuk Perguruan Tinggi Negeri.

Ada beberapa model penskoran pada bentuk tes pilihan ganda yaitu *correct score*, *punishment score* dan *reward punishment score* (Linda & Algina:

1986, 399). Model penskoran pada bentuk tes pilihan ganda yang digunakan oleh sebagian besar tingkat satuan pendidikan adalah dengan cara menjumlahkan jawaban betul pada lembar jawaban peserta tes. Model penskoran demikian dinamakan *correct score*. Pada model penskoran ini memberi peluang peserta tes untuk menebak jawaban yang telah tersedia pada perangkat tes sehingga memiliki dua kemungkinan jawaban peserta tes yaitu benar atau salah. Tebakan *(guessing)* dalam tes pilihan ganda dapat menurunkan nilai validitas butir dan reliabilitas tes (Hopkins & Antes dikutip Santosa: 2011: 135). Model penskoran untuk menghindari sedikit peluang menebak adalah *punishment score*. Model ini menghitungkan jawaban salah yang direspon oleh peserta tes dengan jalan memberikan hukuman dalam bentuk mengurangi skor (Santosa, 2011: 136)

Sebuah butir pilihan ganda yang direspon oleh dua kelompok peserta tes yang berbeda dan memiliki kemampuan yang sama akan menghasilkan estimasi parameter taraf sukar butir, daya beda butir atau faktor kebetulan jawaban betul berbeda. Butir tersebut tidak mengukur aspek yang sama pada dua kelompok peserta tes yang berbeda (Naga, 1992 : 442). Butir itu turut mengukur aspek lain yang seharusnya tidak diukur, sehingga skor dari dua kelompok peserta tes yang seharusnya tidak berbeda menjadi berbeda. Butir demikian dinamakan bias atau *differential item functional* (*DIF*) (Barr dan Raju: 2003). *DIF* merupakan perbedaan yang tak terduga antara dua kelompok peserta tes yang seharusnya sama pada atribut yang diukur dengan item (Dorans&Holland, 1993: 37). Suatu butir menunjukkan *DIF* apabila peserta tes yang memiliki kemampuan yang sama berada dalam kelompok yang berbeda, tidak mempunyai probabilitas yang sama untuk menjawab betul (Barr dan Raju, 2003; Hsin-Hung dan Stot, 1995; Hambleton, Swaminathan, & Rogers, 1991)

Deteksi *DIF* pada suatu butir menggunakan analisis statistika parametrik *diantaranya lord's chi-square test, log linear models, IRT LRT, likelihood ratio test, logistik regression* (Wiberg, 2007: 27). Untuk mendeteksi suatu butir mengandung *DIF* dengan menggunakan metode *Lord's Chi-Square* maka terlebih dahulu parameter butir dari dua kelompok peserta tes yang berbeda disetarakan skornya. Penyetaraan skor berdasarkan teori responsi butir dinamakan metode *equating*. Salah satu metode *equating* adalah metode rerata dan sigma. Hasil penelitian Rahayu (2010) menyatakan metode rerata dan sigma, dan metode kurva karakteristik merupakan metode yang paling akurasi untuk penyetaraan skor pada pendeteksian *DIF*.

Penyetaraan skor parameter butir dari dua kelompok berbeda diperlukan respon peserta tes. Respon peserta tes ini diestimasi parameter butir dan paramater kemampuan peserta tes dari dua kelompok berbeda dapat menggunakan ukuran sampel kecil, sedang dan besar. Pada model logistik satu parameter biasa menggunakan ukuran sampel 250, pada model logistik dua parameter menggunakan ukuran sampel 500, dan pada model logistik tiga parameter menggunakan ukuran sampel 1000. Cohen dan Seock (1998), Candel dan Drasgow (1988), Rahayu (2008) membandingkan metode *linking* dalam

pendektesian *DIF*. Cohen dan Seock menggunakan ukuran sampel 300 dan 600, Candell dan Drasgow menggunakan ukuran sampel 300 dan 500, Rahayu menggunakan ukuran sampel 300 dan 600. Variabel terikat pada penelitian Cohen dan Seock adalah butir *false positive* dan butir *false negative*, variabel terikat pada penelitian Candell dan Drasgrow adalah banyaknya butir *DIF*, variabel terikat pada penelitian Rahayu adalah butir *hit*.

Penelitian yang berkaitan dengan *DIF* pada butir yang berbentuk pilihan ganda dengan model penskoran *correct score* telah banyak dilakukan diantaranya oleh Effendi (2011) dan Sudaryono (2012), Rahayu (2008), sedangkan penelitian yang berkaitan dengan *DIF* dengan memperhatikan model penskoran dan metode *equating* belum ada. Oleh karena itu tujuan penelitian ini untuk menentukan apakah semakin panjang ukuran sampel, maka semakin sensitif deteksi DIF dengan menggunakan metode *Lord Chi-Square* berdasarkan model penskoran *correct score* dan *punishment score* pada model logistik dua parameter.

METODE PENELITIAN

Metode penelitian yang digunakan adalah metode eksperimen. Data yang digunakan dalam penelitian ini berupa skor pekerjaan siswa pada mata pelajaran kimia di Jakarta pada tahun ajaran 2000. Skor hasil pekerjaan siswa ini berbentuk nol (0) dan satu (1). Butir perangkat tes yang cocok model dengan L2P adalah 46 butir.

Tabel 1. Disain Penelitian Deteksi DIF ditinjau Model Penskoran dan Ukuran Sampel

Ukuran Sampel	**Model Penskoran (A)**	
	Correct Score	***Punishment Score***
Kecil	$Y_{1.1.1}$	$Y_{1.2.1}$
	.	.
	.	.
	$Y_{1.1.30}$	$Y_{1.2.30}$
Sedang	$Y_{2.1.1}$	$Y_{2.2.1}$
	.	.
	.	.
	$Y_{2.1.30}$	$Y_{2.2.30}$
Besar	$Y_{3.1.1}$	$Y_{3.2.1}$
	.	.
	.	.
	$Y_{3.1.30}$	$Y_{3.3.30}$

Keterangan Y_{ijk} = banyak butir DIF

Variabel bebas dalam penelitian ini adalah model penskoran yaitu model penskoran *correct score* dan *punishment score* dan ukuran sampel yaitu 300, 500

dan 800. Variabel terikat adalah banyak butir *DIF*. Ukuran sampel 300, 500 dan 800 berturut-turut dinamakan sampel kecil, sampel sedang dan sampel besar. Skor pekerjaan siswa diambil dengan teknik *random sampling* dan pengulangan sebanyak 30 kali.

Equating menggunakan metode rerata dan sigma (RS) hanya melibatkan taraf sukar butir. Untuk penyetaraan skor menggunakan rumus $b^*_{jF} = Ab_{jF} + K$, dengan $A = \frac{\sigma_{b_R}}{\sigma_{b_F}}$ dan $K = \mu_{b_R} - A\mu_{b_F}$. Rumus penyetaraan skor ini mengandung koefisien A dan K. Koefisien A diperoleh dari perbandingan antara simpangan baku taraf sukar butir kelompok referensi dan kelompok fokal, sedangkan koefisien K diperoleh dari perbedaan rerata taraf sukar butir kelompok referensi, dan perkalian A dengan rerata taraf sukar butir kelompok fokal. Estimasi parameter butir kelompok fokal dan referensi dilakukan secara tepisah.

Lord memberikan rumus transformasi linear untuk taraf sukar butir (Cohen dan Seock, 1998) adalah $b^*_{j2} = Ab_{j2} + K$,* menyatakan nilai transformasi, sehingga estimasi parameter taraf sukar butir ke-j pada kelompok dua sama skalanya dengan kelompok satu. Hasil *equating* parameter butir kelompok fokal dipengaruhi oleh nilai A dan K, sedangkan nilai A dan K dipengaruhi oleh nilai taraf sukar butir kelompok fokal dan kelompok referensi.

Stark dan Chernyshenko (2002:7) mengemukakan langkah-langkah yang dilakukan dalam pendektesian *DIF* dengan menggunakan metode *Lord's Chi-Square* lebih jelas dibandingkan dengan yang dikemukan Segal adalah:

1) Mengestimasi parameter butir dan variansi kovariansi kelompok kelompok fokal dan kelompok referensi secara terpisah.
2) Menyamakan skala dari kelompok fokal ke kelompok referensi
3) Menghitung nilai χ^2 dan membandingkan harga kritis χ^2.

Prosedur pelaksanaan tes hasil belajar dengan menerapkan model penskoran adalah sebelum peserta merespon perangkat tes hasil belajar, pengawas menginformasikan bahwa model penskoran yang digunakan adalah *correct score* dan *punishment score*. Respon peserta tes yang digunakan untuk deteksi DIF pada model penskoran *correct score* dan *punishment score* adalah skor dengan perhitungannya sama yaitu benar diberi skor 1 dan salah diberi skor 0.

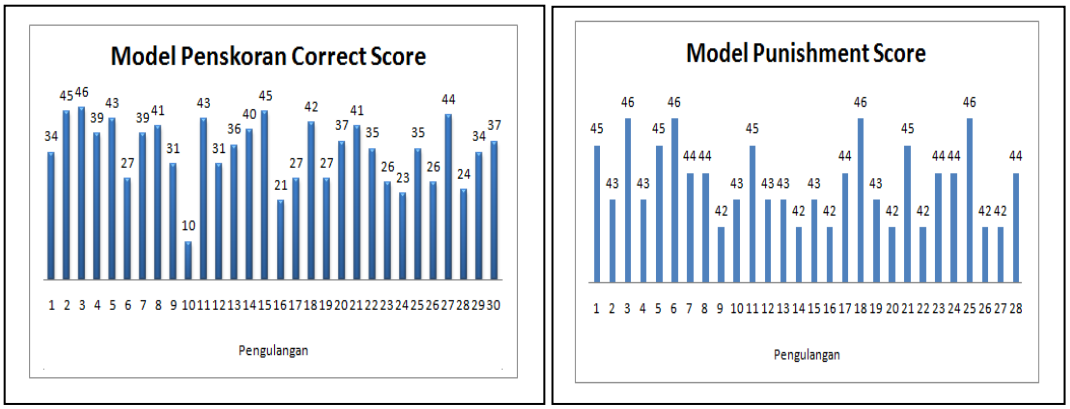
Tahap pendeteksian DIF yang dilakukan yaitu langkah pertama menggandakan respon peserta tes dari wilayah Jakarta Utara dan Jakarta Selatan. Langkah kedua menentukan butir yang cocok model dengan model logistik dua parameter (L2P), langkah ketiga mengambil secara acak 300, 500 dan 800 skor responden untuk kelompok fokal dan kelompok referensi, masing-masing dilakukan sebanyak 30 pengulangan. Langkah keempat, melakukan estimasi parameter taraf sukar butir dan daya beda butir, selanjutnya dilakukan penyetaraan skor dengan metode rerata dan sigma. Langkah kelima deteksi

*DIF*nya dengan metode *Lord Chi-Square*. Langkah keenam melakukan pengujian hipotesis dengan menggunakan anava dua faktor.

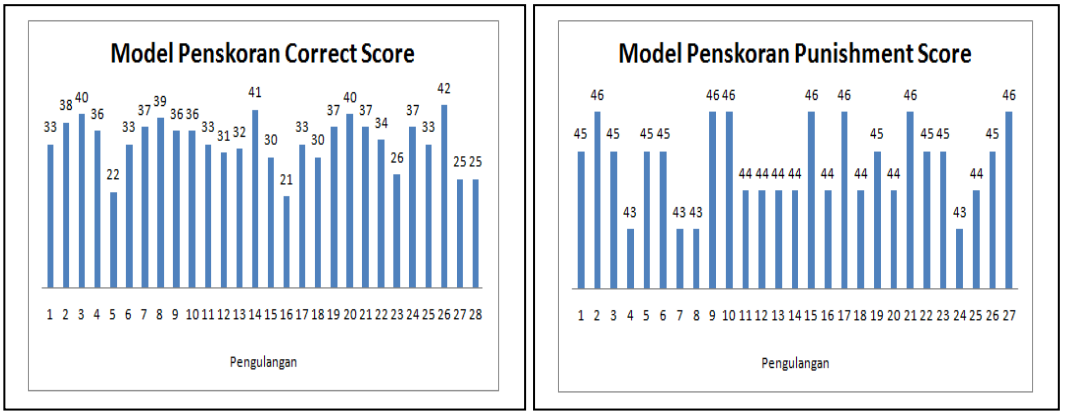
HASIL PENELITIAN

Butir perangkat tes mata pelajaran kimia tahun 2003 yang cocok model dengan model logistik dua parameter terdiri 4 butir tidak cocok model dan 46 butir cocok model. Banyak butir DIF dengan metode *Lord Chi-Square* pada model penskoran *punishment score* dengan ukuran sampel kecil adalah 42 butir sampai dengan 46 butir. Banyak butir DIF yang sering terdeteksi sebanyak 43 butir terjadi 7 pengulangan yakni pengulangan ke 2, 4, 10, 12, 13, 15, dan 19. Banyak butir pada model penskoran *correct score* adalah 10 butir sampai dengan 46 butir. Banyak butir yang terdekteksi pada model penskoran ini lebih beragam dan banyak butir DIF kurang dari 20 butir yaitu sebanyak 10 butir muncul satu kali pengulangan yaitu pada pengulangan ke 10.

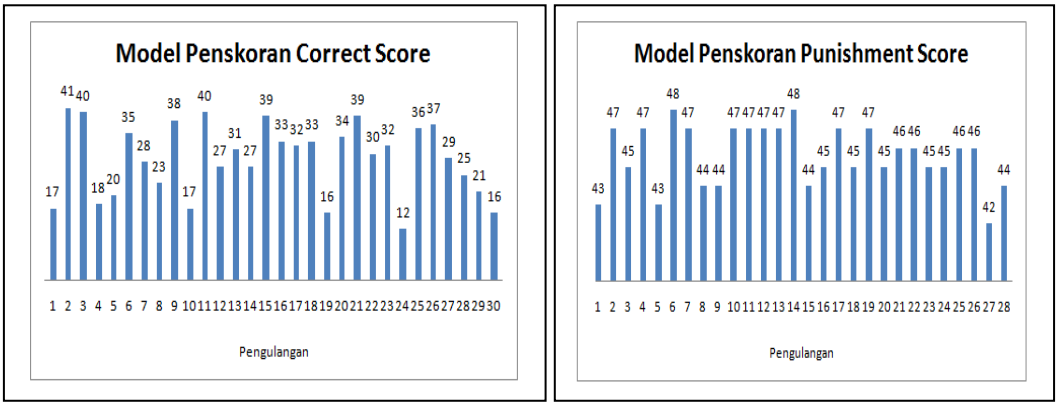
Banyak butir DIF dengan metode *Lord Chi-Square* pada model penskoran *punishment score* dengan ukuran sampel sedang adalah 43 butir sampai dengan 46 butir. Banyak butir DIF yang sering terdeteksi adalah sebanyak 44 butir yakni pada pengulangan 11, 12, 13, 14, 16, 18, 20 dan 25 dan 45 butir pada pengulangan ke 1, 3, 5, 6, 19, 22, 23, dan 26. Banyak butir pada model penskoran *correct score* adalah 21 butir sampai dengan 42 butir. Banyaknya butir yang terdeteksi pada model penskoran ini lebih beragam dan banyaknya butir DIF kurang dari 30 butir yaitu sebanyak lima kali pengulangan yaitu sebanyak 21 butir, 22 butir, 25 butir dan 26 butir pada pengulangan ke-5, 16, 23, 27 dan 28.



Gambar 1. Banyak Butir DIF pada Ukuran Sampel Kecil



Gambar 2. Banyak Butir DIF pada Ukuran Sampel Sedang



Gambar 3. Banyak Butir DIF pada Ukuran Sampel Besar

Banyak butir *DIF* dengan metode *Lord Chi-Square* pada model penskoran *correct score* dengan ukuran sampel besar adalah 12 butir sampai dengan 41 butir. Banyak butir setiap pengulangan berbeda kecuali pada pengulangan 16 dan 18, 15 dan 21, 3 dan 11. Sementara banyak butir pada model penskoran *punishment score* adalah 40 butir sampai dengan 48 butir. Banyak butir yang terdeteksi hanya sekali pengulangan yakni 42 butir.

Hasil pengujian diperoleh pada baris model penskoran nilai sgn. 0,000 < 0.05, ini berarti terdapat perbedaan banyaknya butir *DIF* dengan menggunakan model penskoran *correct score* dan *punishment score*. Rata-rata banyak butir *DIF* dengan menggunakan model penskoran *correct score* 32 butir, model penskoran *punishment score* 44 butir, sehingga dapat disimpulkan banyak butir *DIF* dengan menggunakan model penskoran *punishment score* butir lebih banyak dari model penskoran *correct score*.

Tabel 2. Hasil Pengujian dengan Anava Dua Jalan

Tests of Between-Subjects Effects

Dependent Variable:BanyaknyabutirDIF

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7248.284[a]	5	1449.657	45.252	.000
Intercept	253603.084	1	253603.084	7.916E3	.000
UkuranSampel	125.687	2	62.844	1.962	.144
ModelPenskoran	6637.533	1	6637.533	207.194	.000
UkuranSampel * ModelPenskoran	426.780	2	213.390	6.661	.002
Error	5317.873	166	32.035		
Total	264519.000	172			
Corrected Total	12566.157	171			

a. R Squared = .577 (Adjusted R Squared = .564)

Hasil pengujian banyak butir DIF dengan model penskoran *correct score* antara ukuran sampel kecil, sedang, dan besar diperoleh nilai sign. 0,000 < 0,05 maka ada perbedaan banyak butir DIF dengan model penskoran *correct score* pada ukuran sampel kecil, sedang, dan besar. Pada ukuran sampel kecil dan sedang didapat nilai sign. 0,684 > 0,05 maka dapat disimpulkan tidak ada perbedaan banyaknya butir DIF pada model penskoran *correct score* antara ukuran sampel kecil dan sedang. Pada ukuran sampel kecil dan besar didapat nilai sign. 0,008 < 0,05 maka banyaknya butir DIF pada model penskoran *correct score* dengan ukuran sampel kecil lebih banyak dari ukuran sampel besar. Pada ukuran sampel sedang dan besar didapat nilai sign. 0,027 < 0.05 maka banyaknya butir DIF pada model penskoran *correct score* dengan ukuran sampel sedang lebih banyak dari ukuran sampel besar.

Tabel 3. Hasil Uji Perbandingan Pada Model Penskoran *Correct Score*

(I) Ukuran Sampel	(J) Ukuran Sampel	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
kecil	Sedang	.836	2.049	.684	-3.24	4.91
	besar	5.433*	2.013	.008	1.43	9.44
Sedang	kecil	-.836	2.049	.684	-4.91	3.24
	besar	4.598*	2.049	.027	.52	8.67
besar	kecil	-5.433*	2.013	.008	-9.44	-1.43
	Sedang	-4.598*	2.049	.027	-8.67	-.52

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

Hasil pengujian banyak butir *DIF* dengan model penskoran *punishment score* antara ukuran sampel kecil, sedang, dan besar diperoleh nilai sign. 0,006 < 0,05 maka ada perbedaan rata-rata banyak butir *DIF* dengan model penskoran *punishment score* antara ukuran sampel kecil, sedang, dan besar. Pada ukuran sampel kecil dan sedang didapat nilai sign. 0,022 < 0,05 maka dapat dsimpulkan banyaknya butir *DIF* pada model penskoran *punishment score* dengan ukuran sampel sedang lebih banyak dari ukuran sampel kecil. Pada ukuran sampel kecil dan besar didapat nilai sign. 0,000 < 0,05 maka banyaknya butir *DIF* pada model penskoran *correct score* antara ukuran sampel besar lebih banyak dari ukuran

sampel kecil. Pada ukuran sampel sedang dan besar didapat nilai sign. 0,008 < 0,05 maka banyaknya butir *DIF* pada model penskoran *correct score* dengan ukuran sampel besar lebih banyak dari ukuran sampel sedang.

Tabel 4. Hasil Uji Perbandingan Pada Model Penskoran *Punishment Score*

(I) Ukuran Sampel	(J) Ukuran Sampel	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
kecil	Sedang	-.843*	.361	.022	-1.56	-.12
	besar	-1.854*	.368	.000	-2.59	-1.12
Sedang	kecil	.843*	.361	.022	.12	1.56
	besar	-1.011*	.374	.008	-1.76	-.27
besar	kecil	1.854*	.368	.000	1.12	2.59
	Sedang	1.011*	.374	.008	.27	1.76

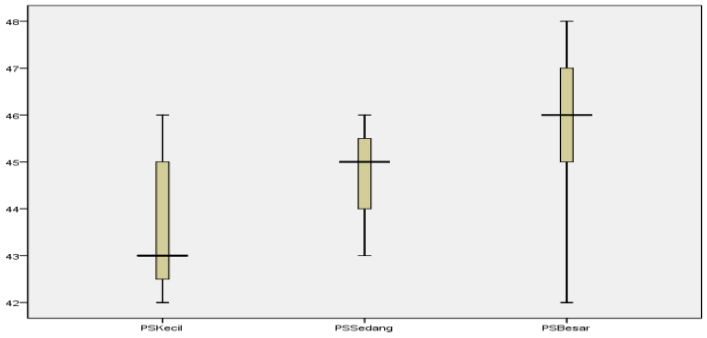
*. The mean difference is significant at the 0.05 level.

PEMBAHASAN

Hasil pengujian pertama, banyaknya butir *DIF* pada model penskoran *correct score* dengan ukuran sampel kecil lebih banyak dari ukuran sampel besar, dan banyaknya butir DIF pada model penskoran *correct score* dengan ukuran sampel sedang lebih banyak dari ukuran sampel besar. Dapat disimpulkan pada ukuran sampel kecil dan sedang lebih banyak terdeteksi butir DIF dengan metode *Lord Chi Square* daripada ukuran sampel besar pada model penskoran *correct score*. Ini sejalan dengan penelitian Rahayu (2008) yaitu dengan model penskoran *correct score,* penyamaan skala dengan metode rerata dan sigma dan deteksi *DIF* dengan *Lord Chi-Square* pada model logistik dua parameter menyatakan bahwa deteksi *DIF* dengan metode *Lord Chi-Square* pada model penskoran *correct score* paling akurat pada ukuran sampel kecil. Temuan ini didukung data secara deskripsi yaitu rata-rata banyaknya butir *DIF* pada ukuran sampel kecil 34 butir dan sedang 33 butir sedangkan pada ukuran sampel besar 38 butir, demikian pula sebaran banyaknya butir DIF yang terdeteksi, ukuran sampel sedang lebih homogen dibandingkan ukuran sampel kecil dan besar.

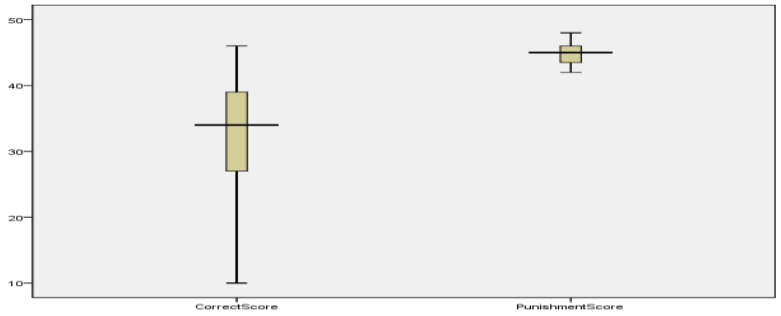
Gambar 4. Boxplot Banyaknya Butir DIF dengan Model *Correct Score* pada Ukuran Sampel Kecil, Sedang dan Besar

Hasil penelitian kedua menunjukkan banyaknya butir *DIF* dengan model penskoran *punishment score* pada ukuran sampel kecil lebih banyak dari ukuran sampel sedang, pada ukuran sampel kecil lebih banyak dari ukuran sampel besar, pada ukuran sampel sedang lebih banyak dari ukuran sampel besar. Sehingga didapat temuan kedua pada penelitian ini adalah deteksi *DIF* dengan metode *Lord Chi-Square* pada model penskoran *punishment score* adalah semakin sedikit ukuran sampel makan makin sedikit butir terdeteksi DIF. Temuan ini didukung data secara deskripsi yaitu rata-rata banyaknya butir DIF pada ukuran sampel kecil dan besar 46 butir, sedangkan pada ukuran sampel besar 48 butir, demikian pula sebaran banyaknya butir DIF yang terdeteksi.



Gambar 5. Boxplot Banyaknya Butir DIF dengan Model *Punishment Score* pada Ukuran Sampel kecil, Sedang dan Besar

Hasil penelitian menunjukkan pada gabungan ukuran sampel kecil, sedang maupun besar adalah banyaknya butir *DIF* dengan model penskoran *punishment score* lebih banyak dari banyaknya butir *DIF* dengan model penskoran *correct score*. Dengan demikian temuan penelitian ini adalah deteksi *DIF* dengan metode *Lord Chi-Square* menghasilkan banyaknya butir *DIF* dengan model penskoran *punishment score* lebih banyak dari banyaknya butir *DIF* dengan model penskoran *correct score* tidak dipengaruhi oleh ukuran sampel. Banyaknya butir DIF yang terdeteksi pada model penskoran *corecct score* dan *punishment score* nampak jelas sekali perbedaannya secara visual pada gambar 6.



Gambar 6. Boxplot Banyaknya Butir DIF dengan Model *Punishment Score* dan *Correct Score*

Perbedaan banyaknya butir *DIF* pada model penskoran *corecct score* dan *punishment score* terjadi disebabkan pada model penskoran *correct score* memberi peluang peserta tes menebak jawaban butir tes sehingga memiliki dua kemungkinan jawaban peserta tes yaitu benar atau salah. Peserta tes akan melakukan tebakan terhadap butir yang tidak ditemukan jawabannya. Model penskoran *correct score* akan akan menyebabkan peserta tes berspekulasi dalam menjawab tes (Santosa, 2011: 135). Penyebab peserta tes melakukan tebakan terhadap jawaban butir tes karena butir tes tersebut terlalu sulit untuk tingkat kemampuannya. Sementara model penskoran *punishment score* memberikan sedikit peluang menebak sehingga skor yang diperoleh melalui model penskoran ini dapat menggambarkan kemampuan peserta tes yang sesungguhnya. Peserta tes akan menjawab butir tes sesuai dengan tingkat kemampuannya.

Faktor banyaknya butir *DIF* pada tes dengan model penskoran *punishment score* adalah kecemasan, dan rasa takut peserta tes terhadap *punishment* yang dilakukan yaitu berupa pengurangan nilai pada butir yang dijawab salah. Santrock menyatakan kecemasan berupa perasaan tidak menyenangkan akan ketakutan (Naswati, 2012: 60). Ini mengakibatkan peserta tes akan menghasilkan respon peserta tes terhadap butir yang diukur dengan perangkat tes. Peserta tes yang merespon instrumen dengan model penskoran *punishment score* juga memiliki sifat kehati-hatian yang lebih tinggi dibandingkan dengan peserta tes dengan model penskoran *correct score.*

Pada model penskoran *punishment score* menunjukkan banyaknya butir DIF lebih banyak dibandingkan dengan model penskoran *correct score*. Ini menunjukkan bahwa dua kelompok peserta yang berbeda yang memiliki kemampuan yang sama merespon sebagian besar butir tes dengan model penskoran *punishment score*, mendapatkan probabilitas jawaban benar yang tidak sama pada butir tersebut. Perserta tes memiliki rasa kecemasan, rasa takut dan kehati-hatian yang tinggi, merespon butir yang tidak sesuai dengan tingkat kemampuannya, sehingga diperoleh butir *DIF* yang lebih banyak dengan menggunakan model penskoran *punishment score.* Banyaknya butir DIF disebabkan kecemasan peserta tes ketika mengerjakan tes, ketidakhati-hatian responden ketika mengerjakan soal tes dan belum terbiasa dengan sistem baru yang digunakan. Naga (1992: 391) menyatakan kecemasan merupakan salah satu penyebab ketidakwajaran skor pada responden. Ini mengakibatkan butir menjadi *DIF.*

SIMPULAN

Semakin besar ukuran sampel maka semakin sedikit butir DIF yang terdeteksi dengan metode *Lord Chi Square* pada model penskoran *correct score* dan semakin kecil ukuran sampel maka semakin banyak butir DIF yang terdeteksi dengan metode *Lord Chi Square* pada model penskoran *correct score.* Penelitian

ini dapat dilanjutkan pada *equating* dua tahap dengan memanipulasi model penskoran dan ukuran sampel dengan menggunakan teori klasik.

DAFTAR PUSTAKA

Barr, Michael A. and Nambury S. Raju. (2003). "IRT-Based Assessments of Rater Effects in Multiple-Source Feedback Instruments." *Organizational Research Methods*, Vol. 6(1).

Candell, G. L., & Drasgow, F. (1988). "An Iterative Procedure for Linking Metrics and Assessing Item Bias in Item Response Theory." *Applied Psychological Measurement*, Vol. 12(3).

Cohen, Allan S. dan Seock-Ho Kim. (1998). *Comparison of Linking and Concurenrent Calibration Under Item Rersponse Theory*. Journal Applied Psychological Measurement. Vol 22(2).

Effendi. (2012). "Detection of Crossing Dif: A Comparison of Raju's Area Measure, Lord's Chi-Square, And Likelihood Ratio Test." *Jurnal Evaluasi Pendidikan*. Vol. 2(2).

Hambleton, Ronald K, H. Swaminathan, dan H. Jane Rogers. (1991). *Fundamentals of Item Response* Theory. London : Sage Publications.

Hsin-Hung, Li. dan William Stout. (2003). *New Prosedure for Detecting of Crossing DIF*. University of Illinois at Urbana-Champaign online. dari http://www. stat.uiuc.edu/paper/Li 941,pdf.

Linda, Crocker dan James Algina. (1986). *Introduction to Clasiical and Modern Test Theory*. Florida: Harcourt Brace Jovanovicch.

Naga, Dali Santun. (1992). *Teori Sekor pada Pengukuran Mental*. Jakarta: Nagarani Citrayasa.

Naswati. (2012). "Kecemasan dalam Ketidakwajaran Skor." *Jurnal Evaluasi Pendidikan*, Vol. 3(1)

Rahayu, Wardani. (2008). "Linking Method and Flase Positive Item on DIF Detection Based on Item Respons Theory." *Jurnal Penelitian dan Evaluasi Pendidikan*, Tahun 14 (1).

Stark, S. dan Oleksandr Chernyshenko. (2002). *Detection of differential item/test functioning (DIF/DTF) Using IRT*. University of Illinois at Urbana Champain online. http://www.work.psych.uiuc.edu/irt/.

Sudaryono. (2012). "Perbandingan Sensitivitas Metode Chi-Square Scheuneman, Mantel-Haenszel, Dan Teori Responsi Butir Model Rasch Pada Pendeteksian *Differential Item Functioning* (*DIF*)." *Jurnal Evaluasi Pendidikan*, Vol. 3(1).

Wiberg, Marie. (2007). *Measuring and Detecting Differential Item Functioning in Criterion-Referenced Licencing Test*. ISSN 1103-2685 Universitas UMEA http://www.edusci.umu.se.