Efek Sub-Sampel pada Deteksi Differential Item Functioning (DIF) dengan Metode Regresi Logistik

Rahmawati
Pusat Penilaian Pendidikan Balitbang
Kementerian Pendidikan dan Kebudayaan RI

Abstrak
Salah satu langkah dalam melakukan studi mengenai differential item function *(DIF) adalah menentukan peserta tes menjadi kelompok referensi (acuan) dan kelompok pembanding. Pada banyak kasus implementasi tes, jumlah sampel antara kelompok referensi dan kelompok pembanding berbeda secara ekstrim. Penelitian ini membandingkan efisiensi dua metode sampling dalam mendeteksi DIF: ukuran sampel setara dan tidak setara. Pada penelitian ini dilakukan studi empiris maupun simulasi. Pada studi empiris, 448 peserta di kelompok pembanding dibandingkan dengan kelompok referensi baik dengan jumlah setara ataupun tidak setara. Studi simulasi mencoba 18 kondisi: 2 ukuran sampel, 3 model DIF, dan 3 distribusi kemampuan. Dengan menggunakan ukuran sampel yang setara, studi empiris mendeteksi 4 dari 57 butir soal berpotensi DIF. Sedangkan perbandingan dengan ukuran sampel tidak setara, tidak mendeteksi adanya butir soal berpotensi DIF. Hasil serupa diperoleh dari studi simulasi. Dari studi ini disimpulkan bahwa melakukan sub sampel pada kelompok referensi merupakan metode yang lebih efisien untuk mendeteksi DIF; jika ukuran sampel berbeda secara ekstrim.*

Kata kunci:
differential item function, kelompok referensi, kelompok pembanding

Pendahuluan

Latar Belakang

Keadilan adalah hal yang penting dalam melaksanakan pengujian. Banyak kebijakan dibuat untuk memastikan setiap peserta mendapat perlakuan yang adil dalam melaksanakan pengujian. Salah satu contoh kebijakan adalah pemberian akomodasi (perlakukan khusus) kepada peserta yang memerlukan, seperti: tambahan waktu, alat bantu khusus, ataupun membacakan soal kepada peserta. Akomodasi tes semacam itu diasumsikan tidak memberikan keun-

Penulis adalah staf pada Pusat Peilaian Pendidikan Balitbang Kemendikbud RI. Korespondensi tentang artikel ini dapat menghubungi: redaksi_jp3i@yahoo.co.id

tungan lebih dan hasilnya setara dengan kelompok yang tidak mendapatkan akomodasi.

Salah satu cara untuk membuktikan kesetaraan hasil tes antara dua kelompok peserta tes adalah dengan melakukan studi *differential item functioning* (DIF). Jodoin dan Gierl (2001) menyebutkan prosedur DIF sebagai metode psikometrik yang dilakukan untuk membuktikan adanya keadilan pada tes. Langkah yang dilakukan dalam studi DIF adalah membagi peserta tes menjadi kelompok referensi dan kelompok pembanding. Kemudian suatu kriteria (umumnya skor tes) kedua kelompok tersebut dibandingkan.

Banyak permasalahan timbul dari penentuan kelompok, misalnya perbedaan distribusi skor antara kelompok referensi dan kelompok lainnya. Jodoin dan Geirl (2001) mempelajari bahwa dengan distribrusi skor yang berbeda maka tingkat kesalahan type I menjadi lebih besar. Pada studi ini permasalahan yang dibahas adalah perbedaan ukuran sampel.

Sebuah tes di Amerika Serikat memberikan akomodasi kepada peserta yang bukan *English native speaker* berupa video yang menayangkan soal dan soal dibacakan keras sesuai dengan penunjuk pada video. Jumlah peserta yang mendapat akomodasi sebesar 448 peserta sedangkan peserta yang menjalani tes dengan prosedur standar sebanyak 117.000 peserta. Perbedaan ukuran sampel antara kedua kelompok sangatlah ekstrim. Terdapat dua alternatif penentuan sampel kelompok: menggunakan ukuran sampel yang setara dan menggunakan ukuran sampel yang sebenarnya.

Rumusan Masalah

Pertanyaan yang dirumuskan pada penelitian ini adalah:

- Bagaimanakah teknik sampel yang efisien untuk mendeteksi DIF dengan metode regresi logistik, jika ukuran sampel dua kelompok berbeda secara ekstrim?
- Bagaimanakah karakteristik teknik sampel tersebut pada bentuk distribusi skor yang berbeda?
- Bagaimanakah karakteristik teknik sampel tersebut saat mendeteksi jenis DIF yang berbeda?

Tujuan Penelitian

Penelitian ini bertujuan untuk:

- Menentukan teknik sampel yang lebih efisien untuk mendeteksi DIF dengan metode regresi logistik.
- Mengetahui karakteristik teknik sampel pada bentuk distribusi skor yang berbeda: distribusi normal, distribusi skew negatif, serta distribusi skew positif.
- Mengetahui karakteristik teknik sampel saat mendeteksi jenis DIF yang berbeda: a, b, dan ab DIF.

Metodologi

Regresi Logistik sebagai Metode Deteksi DIF

Regresi logistik adalah suatu metode yang memodelkan peluang menjawab benar suatu soal berdasarkan „status keanggotaannya". Misalnya kita menentukan ada kelompok laki-laki dan perempuan., maka regresi logistik menghitung peluang menjawab benar suatu soal berdasarkan status jenis kelaminnya. Peluang tersebut dirumuskan dengan:

$$p(u=1|x)=\frac{e^{f(x)}}{1+e^{f(x)}}$$

Dimana $p(u=1|x)$ adalah peluang menjawab benar. Sedangkan f(x) adalah fungsi kombinasi linear dari prediktor (Hidalgo & Lopez-Pina, 2004). f(x) dalam studi DIF dirumuskan sebagai:

$$f(x)=\tau_0+\tau_1\theta+\tau_2 G+\tau_3\theta G$$

Dimana τ_0 adalah intersep, τ_1 adalah koefisien regresi untuk kemampuan (misal skor total tes), τ_2 adalah koefisien regresi untuk variabel pengelompokan (G), dan τ_3 adalah parameter interaksi antara kemampuan dan variabel pengelompokan. Suatu soal dideteksi sebagai potensi uniform DIF jika $\tau_2 \neq 0$ and $\tau_3 = 0$ dan potensi non-uniform DIF jika $\tau_3 \neq 0$. Berdasarkan penelitian Roussos and Stout (1996b), besarnya DIF ditentu-kan oleh perubahan nilai R^2 : kurang 0,035 , DIF dapat diabaikan, antara 0,035 dan 0,07 adalah menengah, sedangkan lebih dari 0,07 dikategorikan besar.

Studi Empiris

Data yang digunakan berasal dari tes matematika kelas 3 suatu state di Amerika Serikat. Tes terdiri atas 57 soal pilihan ganda dan 3 soal uraian. Penelitian ini hanya menganalisis soal pilihan ganda. Deteksi DIF pada studi berdasarkan akomodasi yang diperoleh peserta: peserta yang diperbolehkan melihat teks ujian dalam format video (kelompok pembanding, n = 448) dan peserta yang tidak mendapat akomodasi dan menempuh ujian dalam format cetakan kertas (kelompok referensi, $n \cong 117,000$). Dilakukan sub sampel terhadap kelompok referensi secara *mutually exclusive*, sehingga diperoleh 261 sub-sampel yang terdiri atas 448 peserta. Setiap sub sampel kemudian dibandingkan dengan kelompok akomodasi. Sedangkan kondisi kedua tidak menggunakan sub sampel: seluruh peserta di kelompok referensi dibandingkan dengan kelompok akomodasi.

Untuk mengetahui tingkat *error* tipe I, sebuah sub sampel dari kelompok referensi diperlakukan sebagai *dummy* kelompok akomodasi, kemudian dibandingkan dengan 260 sub-sampel kelompok referensi.

Studi Simulasi

Tiga kondisi dimanipulasi dalam studi simulasi: ukuran sampel, distribusi skor sampel, dan jenis DIF. Di semua kondisi digunakan 50 butir soal dengan proporsi soal DIF sebanyak 20%.

Ukuran sampel dimanipulasi sebanyak 2 kondisi: ukuran setara dan ukuran sampel berbeda secara ekstrim. Data simulasi untuk kelompok referensi sebesar 100.000 peserta, sedangkan kelompok pembanding 500 peserta. Untuk kondisi ukuran sampel setara, 200 sub-sampel berukuran 500 peserta dipilih secara *mutually exclusive* dari kelompok referensi. Sedangkan untuk kondisi ukuran sampel berbeda secara ekstrim tidak dilakukan sub sampel. Selanjutnya dilakukan metode analisis DIF yang sama seperti metode pada studi empiris.

Tabel 1
Nilai Parameter Butir Soal untuk Membangkitkan Data Simulasi Kelompok Referensi

Item	α	β	γ	Item	α	β	γ
1	1.1	-.7	.20	26	.7	.5	.20
2	.7	-.6	.20	27	.7	.5	.20
3	1.4	.1	.20	28	.4	-.4	.20
4	.9	.9	.16	29	.4	-.4	.20
5	**1.2**	**.7**	**.12**	30	**1.2**	**-.5**	**.20**
6	1.6	1.1	.06	31	.7	-1.0	.20
7	1.6	1.1	.06	32	.7	-.2	.20
8	1.6	-.1	.16	33	.7	-.2	.20
9	1.2	.5	.20	34	.5	0.0	.20
10	**2.0**	**1.6**	**.16**	35	**.9**	**.5**	**.14**
11	1.0	1.6	.13	36	1.1	1.4	.04
12	1.5	1.7	.09	37	1.2	-.6	.20
13	1.0	.7	.15	38	1.2	-.6	.20
14	1.1	2.0	.06	39	.6	-.5	.20
15	**1.1**	**2.4**	**.09**	40	**1.6**	**.3**	**.18**
16	2.0	1.4	.11	41	1.1	0.0	.20
17	1.7	1.3	.17	42	1.5	2.0	.06
18	.5	-.6	.20	43	1.9	1.9	.11
19	.9	1.6	.11	44	.9	-.5	.20
20	**1.3**	**.4**	**.18**	45	**.7**	**-.5**	**.20**
21	1.1	1.2	.05	46	1.4	1.6	.11
22	1.2	1.1	.05	47	1.4	1.6	.11
23	1.3	.2	.20	48	1.0	1.7	.08
24	1.3	.2	.20	49	1.2	1.1	.15
25	**.5**	**-.8**	**.20**	50	**1.2**	**1.1**	**.15**

Butir soal yang dicetak tebal merupakan butir soal yang kemudian disimulasi menjadi potensi DIF pada kelompok pembanding.

Bentuk distribusi kemampuan dimanipulasi menjadi tiga kondisi: distribusi normal (mean=0, standar deviasi=1), distribusi *skew* negatif (distribusi beta dengan a=4 b=2), distribusi *skew* positif (distribusi beta dengan a=2 b=4). Sedangkan jenis DIF dimanipulasi menjadi tiga kondisi: a (sebesar -0,5), b (sebesar 0,3), dan ab (sebesar -0,5 dan 0,3) DIF. Total terdapat 18 kondisi manipulasi. Setiap kondisi direplikasi sebanyak 100 kali sehingga untuk setiap kondisi dilakukan analisis terhadap 20.000 perbandingan kelompok ukuran sampel setara dan 100 perbandingan kelompok ukuran tidak setara.

Tingkat *error* tipe 1 ditentukan dengan melihat proporsi soal non-DIF terdeteksi sebagai DIF.

Hasil dan Pembahasan

Studi Empiris

Dengan membandingkan ukuran sampel yang setara, empat butir soal dideteksi sebagai soal DIF dengan pendeteksian sebesar 42% – 97% dari 261 sub-sampel. Sedangkan saat dilakukan perbandingan antara seluruh kelompok referensi dengan kelompok pembanding, tidak ada soal yang dideteksi sebagai potensi DIF. Tingkat *error* tipe 1 yang rendah mengindikasikan bahwa keempat butir soal tersebut memang merupakan soal DIF.

Tabel 2
Items detected as DIF in the empirical study

Item	Equal sample sizes comparison		unequal sample sizes comparison		Type I error	
	% detected as DIF	Average $R^2\ \Delta$	$R^2\ \Delta$	p-value	%detected as DIF	Average $R^2\ \Delta$
16	97.3	0.092	0.000	0.212	4.25	0.010
33	45.0	0.035	0.001	0.000	0.00	0.002
41	42.3	0.033	0.000	0.003	0.00	0.002
50	85.4	0.045	0.001	0.000	0.00	0.004

Studi Simulasi

Faktor Ukuran Sampel

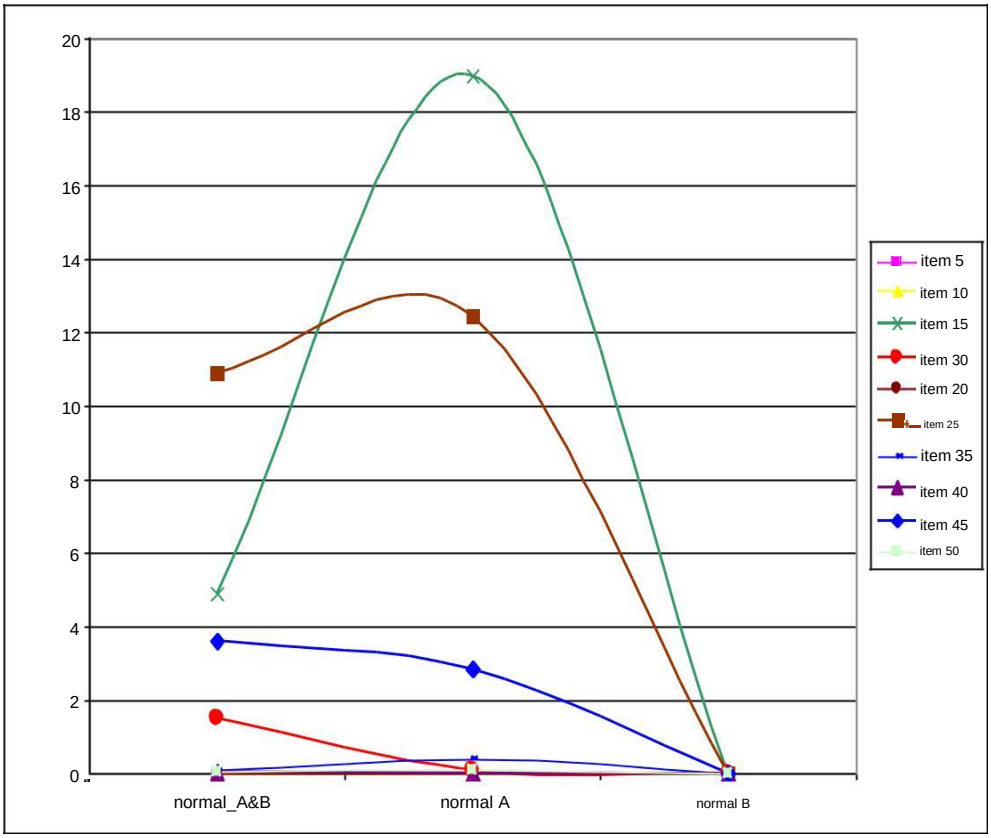
Sesuai dengan hasil studi empiris, untuk kondisi ukuran sampel berbeda secara ekstrim, tidak ada butir soal yang terdeteksi sebagai DIF. Sedangkan pada kondisi ukuran sampel setara, sebanyak 4 soal dari 10 soal terdeteksi sebagai soal DIF.

Tabel 3
Tingkat pendeteksian DIF untuk kondisi ukuran sampel setara

Kondisi	Item 5	Item 10	Item 15	Item 20	Item 25	Item 30	Item 35	Item 40	Item 45	Item 50
normal_A&B	0.01	0.00	4.89	0.01	10.91	1.53	0.08	0.03	3.60	0.04
Positively skewed A&B	0.00	0.00	15.13	0.00	7.66	14.65	0.01	0.05	2.13	0.02
Negatively skewed A&B	0.00	0.00	8.21	0.13	31.17	29.02	0.00	1.33	13.93	0.00
normal A	0.03	0.01	18.99	0.01	12.44	0.09	0.38	0.00	2.86	0.07
Positively skewed A	0.00	0.04	35.60	0.00	11.38	1.23	0.00	0.00	1.64	0.43
Negatively skewed A	0.00	0.12	31.16	0.01	35.42	7.28	0.00	0.01	11.73	0.04
normal B	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.00
Positively skewed B	0.03	0.00	0.01	0.08	0.00	0.05	0.02	0.64	0.00	0.01
Negatively skewed B	0.12	0.02	0.00	0.11	0.00	0.16	0.01	0.49	0.00	0.03

Faktor Jenis Distribusi Kemampuan

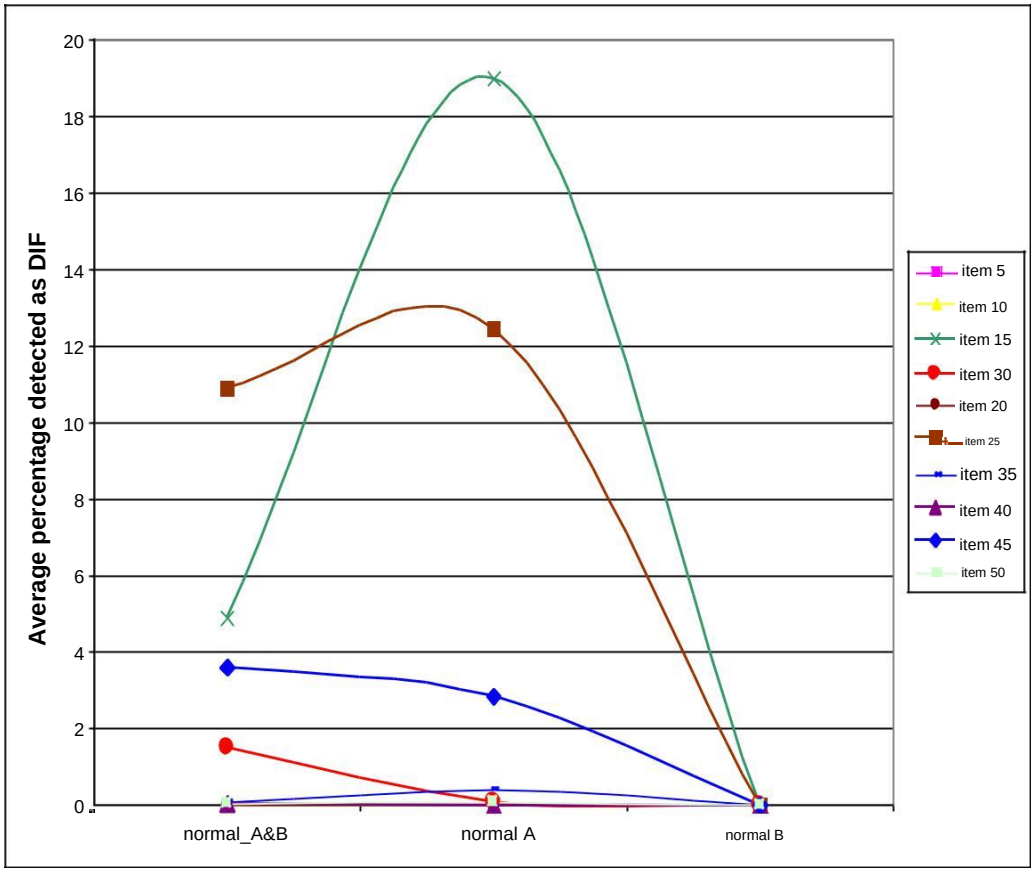
Untuk distribusi normal, tingkat deteksi DIF berkisar antara 0 sampai 19%. Butir soal nomor 15 dan 25 terdeteksi sebagai DIF dengan frekuensi deteksi 12% dan 19%. Untuk distribusi *skew* positif, frekuensi deteksi soal nomor 15 dan 25 lebih besar dari 5%. Sedangkan distribusi *skew* negatif mendeteksi pula soal nomor 30 dan 45 selain soal nomor 15 dan 25. Secara umum, peluang mendeteksi DIF pada bentuk distribusi *skew* lebih besar dibanding-kan bentuk distribusi normal.



Gambar 1. DIF detection rates and type of distribution for a-DIF

Faktor Jenis DIF

Tingkat deteksi DIF lebih tinggi untuk jenis DIF a dibandingkan jenis b dan ab untuk jenis distribusi normal maupun distribusi skew. Sedangkan secara umum, jenis DIF b paling rendah tingkat deteksi DIF-nya baik pada distribusi normal maupun distribusi skew.



Gambar 2. Kaitan antara tingkat deteksi DIF dan jenis DIF pada distribusi normal

Tingkat error tipe 1

Baik pada kondisi ukuran sampel setara maupun tidak setara, tidak satupun butir soal non-DIF terdeteksi sebagai DIF. Tingkat *error* tipe 1 di bawah 0,001. hal ini konsisten dihasilkan di semua kondisi: baik bentuk distribusi, jenis DIF, maupun ukuran sampel.

Kesimpulan dan Saran

Kesimpulan

1. Hasil studi menunjukkan bahwa saat analisis mendeteksi DIF dilakukan dengan menggunakan regresi logistic, ukuran sampel berpengaruh kepada tingkat sensitivitas pendeteksian DIF. Khusus

untuk kasus dimana ukuran sampel antara kelompok referensi dengan kelompok pembanding berbeda secara ekstrim, maka metode sub sampel kelompok sehingga jumlah sampelnya setara, terbukti lebih efisien untuk mendeteksi DIF.

2. Pendeteksian DIF dengan metode regresi logistic dipengaruhi oleh jenis distribusi kelompok. Peluang DIF terdeteksi pada distribusi berbentuk skew lebih besar dibandingkan distribusi bentuk normal.
3. DIF jenis a atau *non-uniform* DIF memiliki peluang lebih besar untuk dideteksi sebagai DIF dibandingkan DIF jenis b dan ab. Sedangkan *uniform* DIF relatif lebih rendah tingkat pendeteksian DIF-nya terhadap jenis DIF yang lain.

Saran

1. Pada studi ini butir soal hanya dimanipulasi dengan satu magnitude DIF. Maka studi lanjutan yang menjadikan variabel magnitude DIF sebagai variabel yang dimanipulasi menjadi wacana yang menarik.
2. Studi ini juga hanya meneliti kondisi saat perbedaan ukuran sampel sangatlah ekstrim. Perlu diketahui sampai batas perbedaan seberapakah metode sub-sampel tidak diperlukan. Oleh karena itu disarankan adanya studi lebih lanjut mengenai seberapa besar perbedaan yang berdampak terhadap efektivitas pendeteksian DIF.
3. Selanjutnya pada studi ini bentuk distribusi antara kelompok referensi dan kelompok pembanding dikondisikan sama. Padahal pada kondisi nyata, bukan hal yang jarang bentuk distribusi antara kedua kelompok berbeda. Oleh karena itu penelitian yang mencoba menggali kondisi tersebut sangat disarankan agar lebih „membumi" dengan kondisi di dunia nyata.

Daftar Pustaka

Hidalgo, M. D. & Lopez-Pina, J. A. (2004). Differential item functioning and effect size: A comparison between logistic regression and Mantel Haenszel procedures. *Educational and Psychological Measurement*, 64, 903-915.

Jordoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error & power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Educa-tion*, 14(4), 329-349.

Kirk, R. E. (1996). Practical Significance: A concept whose time has come. *Educational and Psycho-logical Testing*, 56, 746-759.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum"s three-parameter logistic model. *Educational and Psychological Measurement, 28*, 989-1020.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effect of small sample size and standardized item parameters on SIBTEST and Mantel-Haenszel Type I errorperformance. *Journal of Educational Measurement*, 33, 215-230.