Cerdika: Jurnal Ilmiah Indonesia, 6 (7), July 2026
p-ISSN: 2774-6291 e-ISSN: 2774-6534
Available online at http://cerdika.publikasiindonesia.id/index.php/cerdika/index

Kajian Metode Deteksi Differential Item Functioning pada Butir Soal Dimensi Kepribadian

Arien Citha Utami[1]*, Sri Kustiara[2], Anang Kurnia[3], Alona Dwinata[4]
IPB University, Indonesia[1,2,3]
Universitas Maritim Raja Ali Haji, Indonesia[4]
Email: cithautami@gmail.com[1]*

Keywords:
Graded Response Model; Differential Item Functioning; Standardization; Mantel Haenszel.

Abstract
This study aims to detect Differential Item Functioning (DIF) on personality dimension items using the Standardization and Mantel Haenszel methods. Psychological measurements often contain bias when administered to groups or individuals in the form of questions. This bias has a detrimental impact because an individual's abilities can be influenced by other factors such as gender, residence, religion, ethnicity, and education level. There are many factors that form the basis of the problem as an influence on a person's psychological measurement, resulting in discrepancies. This study began by using a Graded Response Model to model data from five groups of items, then conducted Differential Item Functioning (DIF) detection. In this study, the DIF detection methods used were the standardization method and the Mantel Haenszel method. The results of the item group model showed that the Mantel Haenszel method was more sensitive than the standardization method. The results of DIF detection with the Mantel Haenszel method focused on the odds ratio value. This was proven by the similar model character to the identification of Differential Item Functioning (DIF) by looking at the trace plot results.

Kata Kunci:
Graded Response Model; Differential Item Functioning; Standardisasi; Mantel-Haenszel.

Abstrak
Penelitian ini bertujuan untuk melakukan deteksi Differential Item Functioning (DIF) pada butir soal dimensi kepribadian menggunakan metode Standardization dan Mantel Haenszel. Pengukuran psikologis sering kali mengandung bias ketika dilakukan pada kelompok atau individu dalam bentuk pertanyaan. Bias ini memiliki dampak yang merugikan karena kemampuan individu dapat dipengaruhi oleh faktor-faktor lain seperti jenis kelamin, tempat tinggal, agama, etnis, dan tingkat pendidikan. Ada banyak faktor yang menjadi dasar permasalahan sebagai pengaruh pengukuran psikologis seseorang, sehingga menghasilkan ketidaksesuaian. Penelitian ini dimulai dengan menggunakan Graded Response Model untuk memodelkan data dari lima kelompok item, kemudian melakukan deteksi Differential Item Functioning (DIF). Dalam penelitian ini, metode deteksi DIF yang digunakan adalah metode standardisasi dan metode Mantel Haenszel. Hasil model kelompok item menunjukkan bahwa metode Mantel Haenszel lebih sensitif dibandingkan metode standardization. Hasil deteksi DIF dengan metode Mantel Haenszel memiliki fokus pada nilai odds ratio. Dengan dibuktikannya karakter model yang serupa dengan pengidentifikasian Differential Item Functioning (DIF) dengan melihat hasil trace plot.

PENDAHULUAN

Istilah “psikometrika” didefinisikan sebagai ilmu pengukuran psikologis dimana mengusahakan nilai pengujian sesuai dengan kualitas teknisnya. Saat berbicara mengenai psikometrika, penting sekali untuk mengacu pada seberapa konsisten dan seberapa akurat pengukuran uji psikologi. Pengujian digunakan untuk nilai praktis dengan tujuan tertentu (Swerdlik, 2010).

Evaluasi bias pengukuran sangat penting dalam menentukan penilaian pada berbagai kelompok. Salahsatu metode evaluasi bias adalah Differential Item Functioning (DIF), terjadi dengan cara menilai dari bias tingkat item seluruh spesifikasi kelompok dengan membandingkan tanggapan tingkat item antara kelompok yang memiliki skor keseluruhan yang sama. Deteksi Differential Item Functioning (DIF) bisa dilakukan pada kelompok usia, jenis kelamin, tingkat pendidikan, etnis, status kognitif, dan uji bahasa (Garcia et al., 2021).

Dalam konseptual butir soal terindikasi suatu DIF jika diperoleh indikasi yang terletak pada fungsi respons butir atau Item Response Function dengan ditunjukkan melalui kurva karakteristik. DIF menunjukkan bahwa suatu kelompok memiliki skor yang lebih besar dari kelompok lainnya, namun menunjukkan suatu item memiliki tingkat kesulitan berbeda pada grup yang berbeda sehingga melanggar asumsi invariasi parameter (Ajeng, 2022).

Deteksi DIF dengan metode Standardization dikemukakan oleh (Rustam, 2019) butir soal yang terindikasi DIF adalah kemampuan peserta tes yang sama atau skor yang sama, tetapi berbeda dalam menanggapi butir soal. Secara singkatnya dijelaskan dalam skor yang sama, proporsi kelompok referensi dan kelompok focus yang benar dihitung (Agresti, 2018).

Berdasarkan uraian penelitian terdahulu, terdapat kesenjangan penelitian (research gap) yang perlu diisi. Pertama, sebagian besar penelitian DIF pada instrumen kepribadian berfokus pada perbandingan kelompok berdasarkan jenis kelamin atau status pekerjaan, namun belum banyak yang membandingkan deteksi DIF berdasarkan karakteristik demografis lainnya seperti tempat tinggal pada konteks dimensi kepribadian yang beragam . Kedua, penelitian sebelumnya cenderung menggunakan satu metode deteksi DIF saja, tanpa membandingkan performa metode yang berbeda pada data polikotom dengan karakteristik yang bervariasi . Ketiga, penelitian yang mengintegrasikan Graded Response Model dengan perbandingan metode Standardization dan Mantel Haenszel pada data dimensi kepribadian dengan lima kelompok item yang berbeda masih terbatas.

Urgensi penelitian ini didasarkan pada pentingnya identifikasi bias item dalam instrumen pengukuran psikologis untuk memastikan keadilan dan akurasi pengukuran. Bias dalam butir soal dapat menyebabkan kesalahan interpretasi tentang kemampuan atau karakteristik individu, yang berdampak pada keputusan yang diambil berdasarkan hasil pengukuran tersebut. Dalam konteks tes kepribadian, DIF dapat mengakibatkan perbedaan skor yang tidak mencerminkan perbedaan trait yang sebenarnya, melainkan artefak dari karakteristik item yang bias terhadap kelompok tertentu.

Kebaruan penelitian ini terletak pada beberapa aspek. Pertama, penelitian ini menggunakan data dari lima kelompok item dimensi kepribadian yang berbeda (music preferences, movie preferences, hobbies & interests, phobias, spending habits) dengan karakteristik yang bervariasi. Kedua, penelitian ini membandingkan dua metode deteksi DIF (Standardization dan Mantel Haenszel) secara simultan pada data yang sama. Ketiga, penelitian

ini menganalisis hubungan antara hasil deteksi DIF dengan karakteristik model melalui trace plot dari Graded Response Model.

Penelitian ini bertujuan untuk: (1) melakukan deteksi Differential Item Functioning (DIF) pada butir soal dimensi kepribadian menggunakan metode Standardization dan Mantel Haenszel; (2) membandingkan sensitivitas kedua metode dalam mendeteksi DIF; dan (3) menganalisis karakteristik item yang teridentifikasi mengandung DIF melalui trace plot. Penelitian ini diharapkan memberikan kontribusi teoretis dalam pengembangan metode deteksi DIF pada data polikotom, serta manfaat praktis bagi pengembang instrumen psikologis dalam memilih metode deteksi DIF yang tepat berdasarkan karakteristik data dan konteks pengukuran.

Penelitian menunjukkan bahwa pada studi kasus ini akan dibahas perbandingan dua metode deteksi Differential Item Functioning (DIF) menggunakan Standardization dan Mantel Haenszel.

METODE PENELITIAN

Data

Data yang dipakai adalah data sekunder yang didapatkan dari link https://www.kaggle.com/datasets/miroslavsabo/young-people-survey. Data ini adalah data pada tahun 2013 dimana mahasiswa statistika *Faculty of Social and Economic Science* di *Comenius University Bratislava UK* melakukan survei dengan responden teman sekitar. Data terdiri dari 1010 responden penelitian. Survei disajikan kepada peserta dalam bentuk elektronik dan tertulis. Responden berkebangsaan Slovakia dengan rentang usia 15 - 30 tahun. Data merupakan data polikotom dengan berbagai jenis karakteristik kuisioner. Analisis data dilakukan pada beberapa kelompok item dimensi kepribadian sebagai berikut:

1. *Music preferences* (19 item)
2. *Movie preferences* (12 item)
3. *Hobbies & interests* (32 item)
4. *Phobias* (10 item)
5. *Spending habits* (7 item).

Alat

Penelitian ini dilakukan menggunakan dua *software* utama yaitu RStudio dan Microsoft Excel 2019. *Software* RStudio digunakan untuk melakukan deteksi *Differential Item Functioning (DIF)*, sedangkan *software* Microsoft Excel 2019 digunakan untuk menghitung nilai profil demografis dari data.

Prosedur Analisis Data

Penelitian ini dilakukan dengan langkah-langkah sebagai berikut:

1. Melakukan eksplorasi profil demografis pada data survei mahasiswa statistika *Faculty of Social and Economic Science di Comenius University Bratislava UK* dan melakukan *data cleaning* dengan teknik *random forest imputation*.
2. Melakukan analisis data kelompok item menggunakan *Graded Response Model (GRM).*
3. Melakukan deteksi *Differential Item Functioning (DIF)* menggunakan metode *standardization* dan *Mantel Haenszel*.

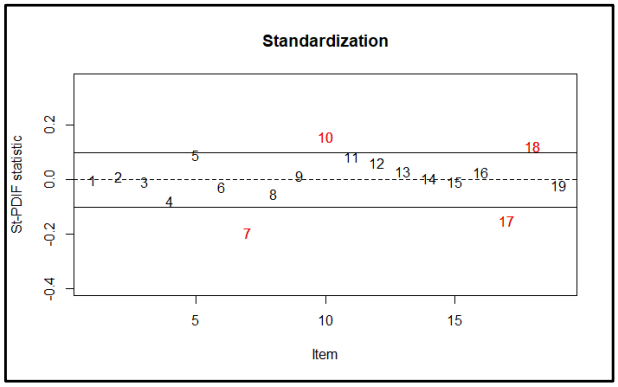
4. Membandingkan hasil deteksi *Differential Item Functioning (DIF)* kedua metode dan melakukan analisis hubungan dengan model.

rumus statistika yang digunakan sebagai bagian dari metode penelitian, sebaiknya tidak menuliskan rumus yang sudah berlaku umum.

HASIL DAN PEMBAHASAN

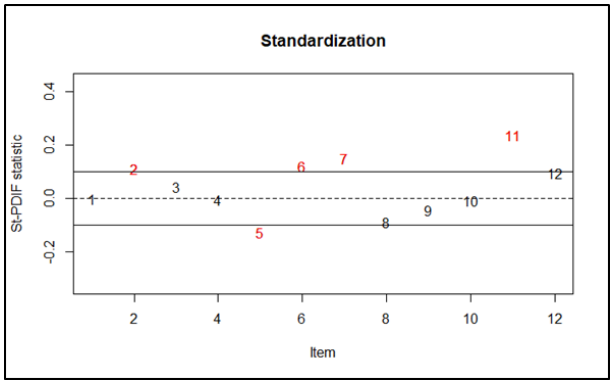
Metode Standardization

Hasil deteksi Differential Item Functioning dengan menggunakan metode standardization ditunjukkan dengan plot item dan penandaan warna merah pada nomor item. Metode ini dilakukan menggunakan dua kriteria profil demografis yaitu jenis kelamin dan tempat tinggal.



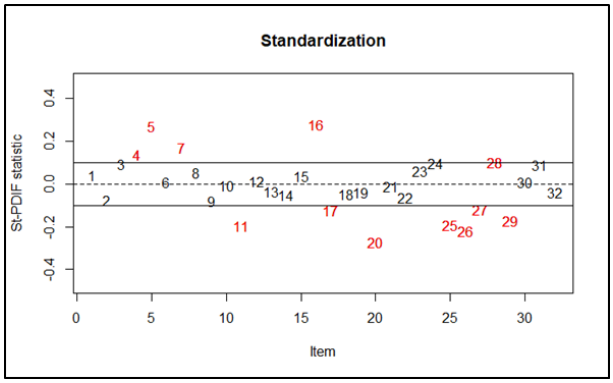
Gambar 1. Deteksi Differential Item Functioning (DIF) menggunakan metode standardization pada kelompok item music preferences berdasarkan jenis kelamin

Gambar 1 menjelaskan bahwa dari 19 jenis music preferences, terdapat 4 item yang terdeteksi memiliki Differential Item Functioning (DIF) yaitu item 7, 10, 17 dan 18. Item tersebut berturut-turut adalah musical, metal/hardrock, latino, dan techno/trance.



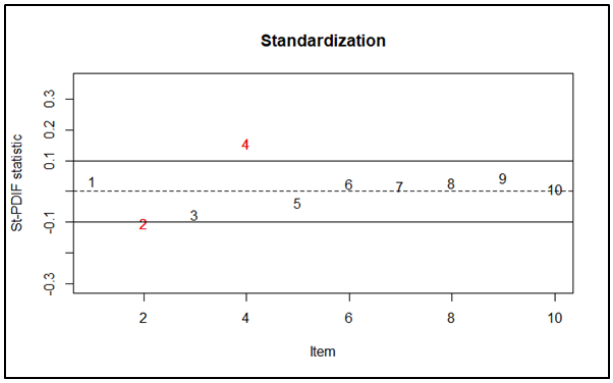
Gambar 2. Deteksi Differential Item Functioning (DIF) menggunakan metode standardization pada kelompok item movie preferences berdasarkan jenis kelamin

Gambar 2 menjelaskan bahwa dari 12 jenis movie preferences, terdapat 5 item yang terdeteksi memiliki Differential Item Functioning (DIF) yaitu item 2, 5, 6, 7, dan 11. Item tersebut berturut-turut adalah horror, romantic, sci-fi, war, dan western.



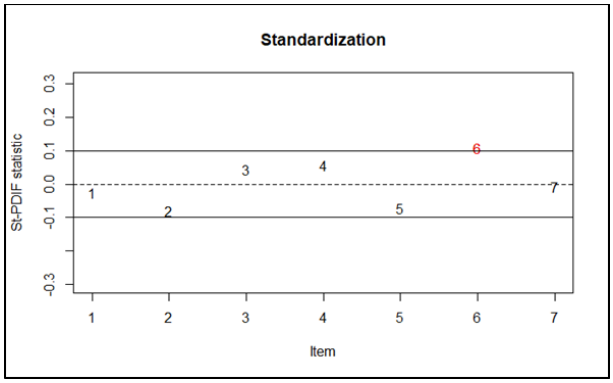
Gambar 3. Deteksi Differential Item Functioning (DIF) menggunakan metode standardization pada kelompok item hobbies and interests berdasarkan jenis kelamin

Gambar 3 menjelaskan bahwa dari 32 jenis hobbies and interests, terdapat 12 item yang terdeteksi memiliki Differential Item Functioning (DIF) yaitu item 4, 5, 7, 11, 16, 17, 20, 25, 26, 27, 28, dan 29. Item tersebut berturut-turut adalah mathematics, physics, PC, reading, cars, art exhibition, dancing, gardening, celebrities, shopping, science and technology, dan theatre.



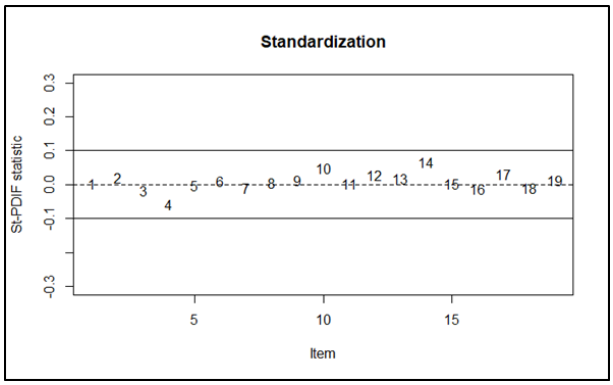
Gambar 4. Deteksi Differential Item Functioning (DIF) menggunakan metode standardization pada kelompok item phobias berdasarkan jenis kelamin

Gambar 4 menjelaskan bahwa dari 10 jenis phobias, terdapat 2 item yang terdeteksi memiliki Differential Item Functioning (DIF) yaitu item 2 dan 4. Item tersebut berturut-turut adalah storm dan heights.



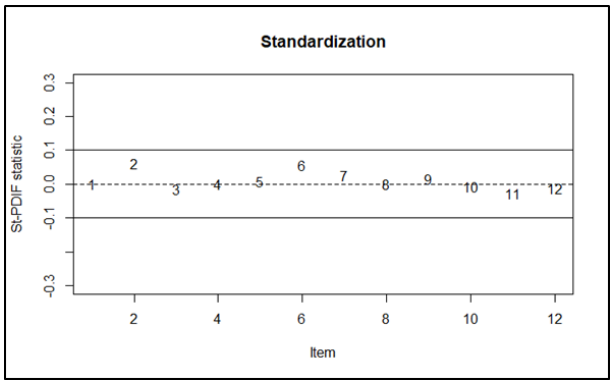
Gambar 5. Deteksi Differential Item Functioning (DIF) menggunakan metode standardization pada kelompok item spending habits berdasarkan jenis kelamin

Gambar 5 menjelaskan bahwa dari 7 jenis spending habits, terdapat 1 item yang terdeteksi memiliki Differential Item Functioning (DIF) yaitu item 6 (spending on gadgets).



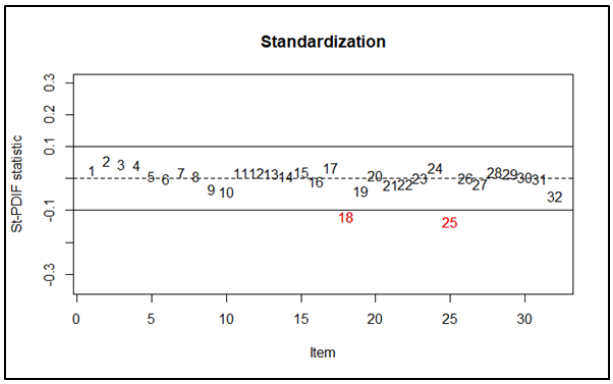
Gambar 6. Deteksi Differential Item Functioning (DIF) menggunakan metode standardization pada kelompok item music preferences berdasarkan tempat tinggal

Gambar 6 menjelaskan bahwa dari 19 jenis music preferences, tidak terdapat sama sekali Differential Item Functioning (DIF).



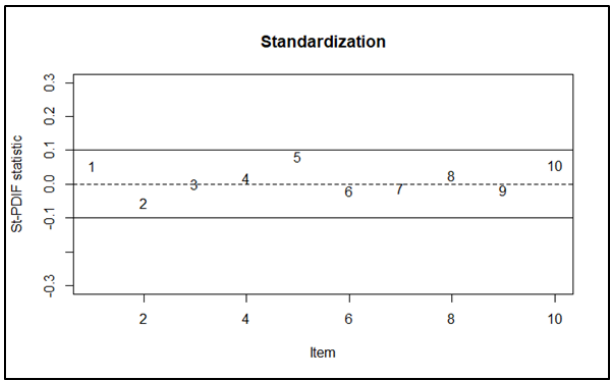
Gambar 7. Deteksi Differential Item Functioning (DIF) menggunakan metode standardization pada kelompok item movie preferences berdasarkan tempat tinggal

Gambar 7 menjelaskan bahwa dari 12 jenis movie preferences, tidak terdapat sama sekali Differential Item Functioning (DIF).



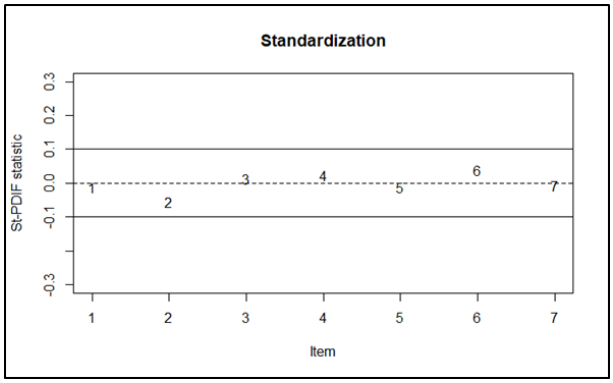
Gambar 8. Deteksi Differential Item Functioning (DIF) menggunakan metode standardization pada kelompok item hobbies and interests berdasarkan tempat tinggal

Gambar 8 menjelaskan bahwa dari 32 jenis hobbies and interests, terdapat 2 item yang terdeteksi memiliki Differential Item Functioning (DIF) yaitu item religion dan gardening.



Gambar 9. Deteksi Differential Item Functioning (DIF) menggunakan metode standardization pada kelompok item phobias berdasarkan tempat tinggal

Gambar 9 menjelaskan bahwa dari 10 jenis phobias, tidak terdapat sama sekali Differential Item Functioning (DIF).

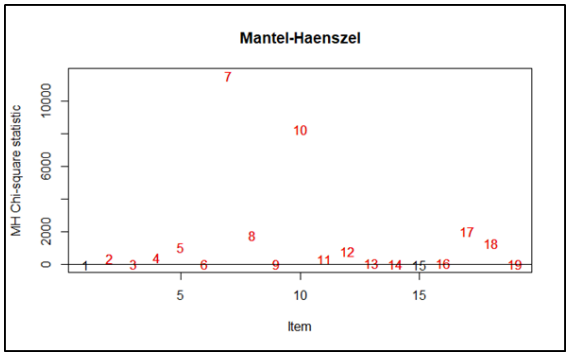


Gambar 10. Deteksi Differential Item Functioning (DIF) menggunakan metode standardization pada kelompok item spending habits berdasarkan tempat tinggal

Gambar 10 menjelaskan bahwa dari 7 jenis spending habits, tidak terdapat sama sekali Differential Item Functioning (DIF).

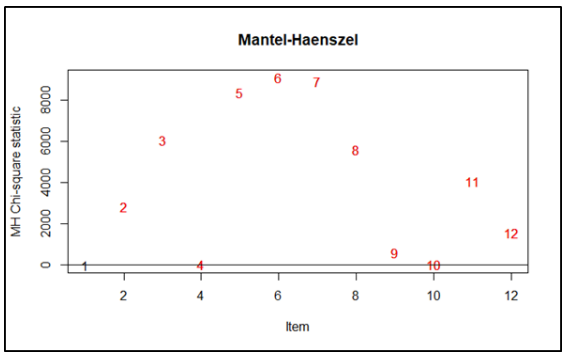
Metode Mantel Haenszel

Hasil deteksi Differential Item Functioning (DIF) dengan menggunakan metode Mantel Haenszel ditunjukkan dengan plot item dan penandaan warna merah pada nomor item. Metode ini dilakukan menggunakan dua kriteria profil demografis yaitu jenis kelamin dan tempat tinggal. Kelompok item phobias dan spending habits tidak ditampilkan dalam plot dikarenakan menghasilkan nilai infinity. Artinya, semua item di kelompok tersebut terdeteksi memiliki Differential Item Functioning (DIF).



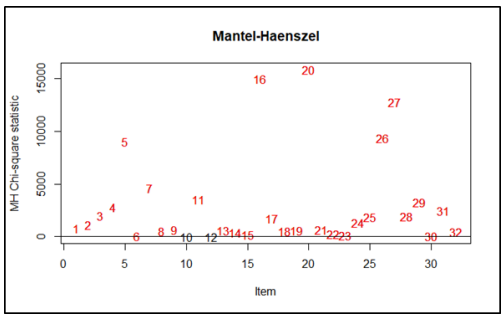
Gambar 11. Deteksi Differential Item Functioning (DIF) menggunakan metode Mantel Haenszel pada kelompok item music preferences berdasarkan jenis kelamin

Gambar 11 menjelaskan bahwa dari 19 jenis music preferences, hampir seluruh item terdeteksi memiliki Differential Item Functioning (DIF), kecuali item 15 yaitu rock and roll.



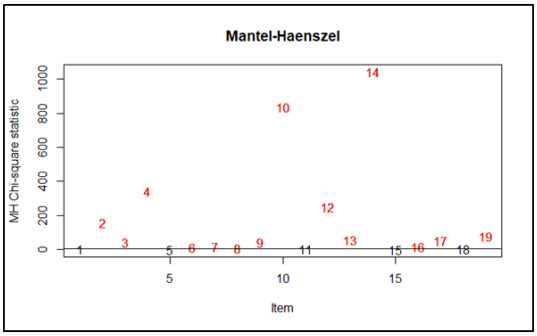
Gambar 12. Deteksi Differential Item Functioning (DIF) menggunakan metode Mantel Haenszel pada kelompok item movie preferences berdasarkan jenis kelamin

Gambar 12 menjelaskan bahwa dari 12 jenis movie preferences, seluruh item terdeteksi memiliki Differential Item Functioning (DIF), dimana item movies tidak diperhatikan karena menilai secara keseluruhan.



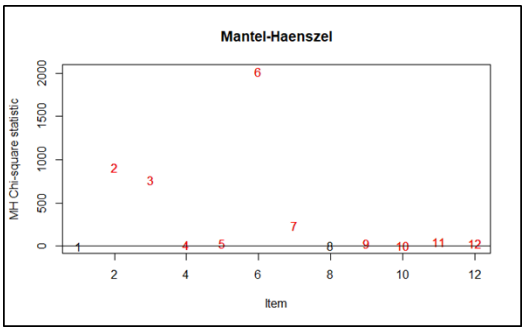
Gambar 13. Deteksi Differential Item Functioning (DIF) menggunakan metode Mantel Haenszel pada kelompok item hobbies and interests berdasarkan jenis kelamin

Gambar 13 menjelaskan bahwa dari 32 jenis hobbies and interests, hampir seluruh item terdeteksi memiliki Differential Item Functioning (DIF), kecuali item 10 dan 12 yaitu chemistry dan geography.



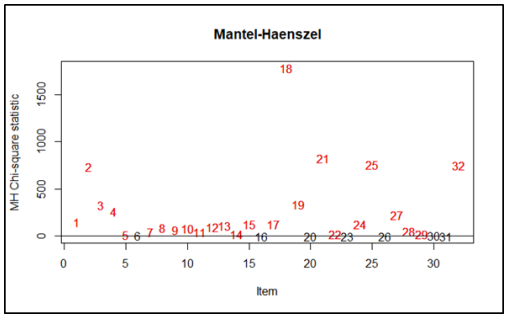
Gambar 14. Deteksi Differential Item Functioning (DIF) menggunakan metode Mantel Haenszel pada kelompok item hobbies and interests berdasarkan tempat tinggal

Gambar 14 menjelaskan bahwa dari 19 jenis music preferences, hampir seluruh item terdeteksi memiliki Differential Item Functioning (DIF), kecuali item 5, 11, 15, dan 18 yaitu country, punk, rock and roll, dan techno/trance.



Gambar 15. Deteksi Differential Item Functioning (DIF) menggunakan metode Mantel Haenszel pada kelompok item phobias berdasarkan tempat tinggal

Gambar 15 menjelaskan bahwa dari 12 jenis movie preferences, hampir seluruh item terdeteksi memiliki Differential Item Functioning (DIF), kecuali item 8 dan 10 yaitu fantasi/fairy tales dan documentary.



Gambar 16. Deteksi Differential Item Functioning (DIF) menggunakan metode Mantel Haenszel pada kelompok item spending habits berdasarkan tempat tinggal

Gambar 16 menjelaskan bahwa dari 32 jenis hobbies and interests, hampir seluruh item terdeteksi memiliki Differential Item Functioning (DIF), kecuali item 6, 16, 20, 23, 26, 30, dan 31. Item tersebut berturut-turut adalah PC, cars, dancing, passive sport, celebrities, fun with friend, dan adrenaline sport.

Perbandingan Metode Deteksi Differential Item Functioning (DIF)

Setelah dilakukan deteksi Differential Item Functioning (DIF), kemudian dilakukan perbandingan terhadap dua metode tersebut.

Tabel 1. Perbandingan metode deteksi Differential Item Functioning (DIF) berdasarkan jenis kelamin

Kelompok item	Jumlah item	*Standardizatin*		*Mantel Haenszel*	
		Terdeteksi	Tidak terdeteksi	Terdeteksi	Tidak terdeteksi
Music preferences	19	4	15	17	2
Movie preferences	12	5	7	12	0
	32	12	20	32	0
Phobias	10	2	8	Inf	0
Spending habits	7	1	6	Inf	0

Tabel 1 menjelaskan bahwa metode Mantel Haenszel dinilai lebih bisa mendeteksi Differential Item Functioning (DIF) pada kelima kelompok item jika dibandingkan dengan metode standardization.

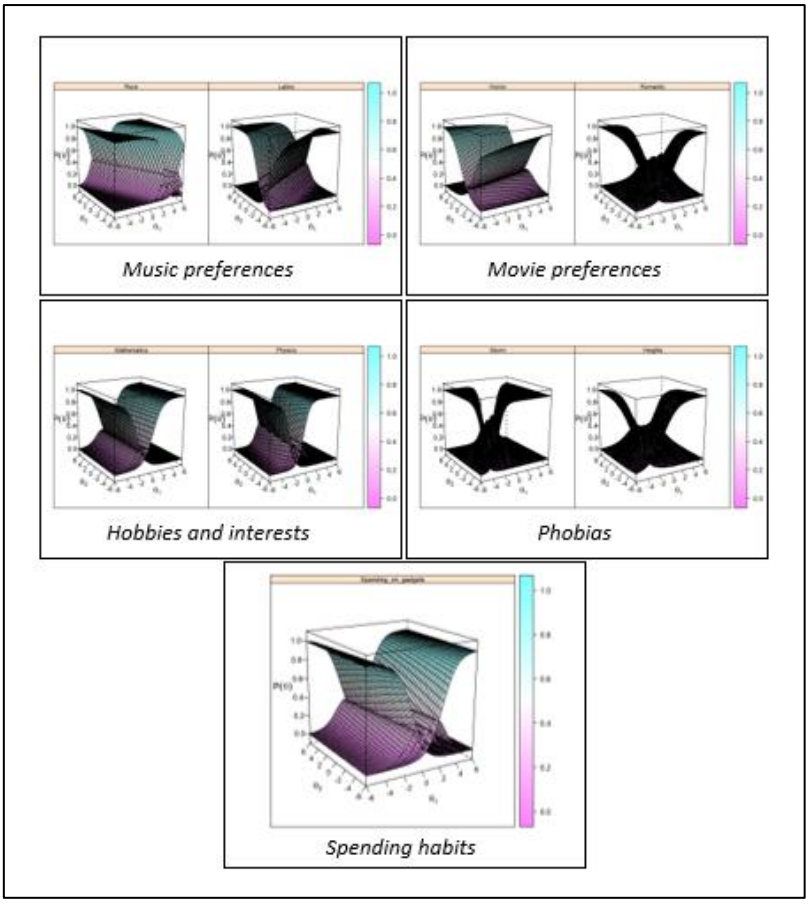
Tabel 2. Perbandingan metode deteksi Differential Item Functioning (DIF) berdasarkan tempat tinggal

Kelompok item	Jumlah item	*Standardization*		*Mantel Haenszel*	
		Terdeteksi	Tidak terdeteksi	Terdeteksi	Tidak terdeteksi
Music preferences	19	0	19	14	5
Movie preferences	12	0	12	10	2
Hobbies & interests	32	2	30	25	7
Phobias	10	0	10	Inf	0
Spending habits	7	0	7	Inf	0

Analisis Lanjutan Model dan Deteksi *Differential Item Functioning (DIF)*

Analisis lanjutan dilakukan pada deteksi *Differential Item Functioning (DIF)* yang dihubungkan dengan *trace* plot 2PL *Graded Response Model (GRM).* Hal ini dilakukan untuk melihat karakteristik item yang teridentifikasi mengandung *Differential Item Functioning (DIF). Trace* plot ini dilakukan kepada 5 kelompok item dengan keterangan item sebagai berikut:

1. *Music preferences (rock, latino)*
2. *Movie preferences (horror, romantic)*
3. *Hobbies and interests (mathematics, physics)*
4. *Phobias (storm, heights)*
5. *Spending habits (spending on gadgets).*



Gambar 17. menjelaskan bahwa dari kelima item yang terdiri dari masing-masing 2 item (kecuali kelompok item spending habits yang hanya 1 item) memiliki karakteristik model yang mirip yaitu trace plot menyerupai huruf "x".

KESIMPULAN

Penelitian ini menyimpulkan bahwa metode Mantel Haenszel lebih sensitif dibandingkan metode standardization dalam deteksi Differential Item Functioning (DIF) beberapa kelompok item. Penelitian terdahulu menyebutkan bahwa metode standardization lebih sensitif terhadap data berukuran kecil, sehingga dikarenakan data yang digunakan mencapai 1010 responden atau dinilai besar maka metode Mantel Haenszel lebih efektif. Hal ini juga disebabkan metode standardization fokus terhadap proporsi kebenaran, sedangkan metode Mantel Haenszel fokus terhadap nilai odds ratio. Item yang teridentifikasi memiliki Differential Item Functioning (DIF) memiliki karakteristik model yang mirip pada datanya dengan melihat trace plot yang dihasilkan.

DAFTAR PUSTAKA

Agresti, A. (2018). *Categorical Data Analysis*.

Ayala RJ (2009). The Theory and Practice of Item Response Theory. New York (USA): The Guilford Press.

Budiyono (2005). Comparison of the Mantel-Haenszel Method, SIBTEST, Logistic Regression and Differences in Chances of Detecting the Presence of DIF. Dissertation, Yogyakarta: Yogyakarta State University.

Djaali (2008). Likert Scale. Jakarta: Main Library

Embretson SE, Reise SP. (2000). Item Response Theory. New York (USA): Psychology Press.
Garcia JM, Gallagher MW, O'Bryant SE, Medina LD (2021). Differential Item Functioning of The Beck Anxiety Inventory in a Rural Multi Ethnic Cohort. Journal of Affective Disorders. 293: 36-42. https://doi.10.1016/j.jad.2021.06.005
French BF, Finch WH, Immekus JC (2019). Multilevel Generalized Mantel Haenszel for Differential Item Functioning Detection. Frontier in Education. 4: 47. https://doi.10.3389/feduc.2019.00047
Mantel, N.,dan Haenszel, W (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. Journal of the National Cancer Institute. 22(4): 719-748.
Haenszel S. Kim, dan Sato A. Kohen (1995). A Comparison of Lord's Chi Square, Raju's Area Measures, and the Likelihood Ratio Test on Detection of Differential Item Function, Journal of Aplied Measurement in Education 8 (1995): 291-312.
Hambleton, Ronald K, H. Swaminathan, dan Rogers, H. J (1991). Fundamentals of Item Response Theory. California: Sage Publications.
Kaplan RM, Saccuzzo DP (2008). Psychological Testing: Principles, Applications, and Issues. 9th ed. Boston (USA): Wadswort Cencage Learning.
Lindberg SM, Hyde JS, Petersen JL, Linn MC (2010). New Trends in Gender and Mathematics Performance: A Meta Analysis. Psychological Bulletin. 136(6): 1123–1135. https://doi.10.1037/a002127
Millsap RE, Kwok OM (2004). Evaluating the Impact of Differential Item Functions on Cross National Comparisons of Test Score Performance. Journal of Educational Measurement. 41(2): 115-139.
Paek I, Cole K (2020). Using R for Item Response Theory Model Applications. New York (USA): Taylor & Francis Group.
Rahayu, W (2008). The Effect of the Linking Method on Many False Positive Items on DIF Detection Based on Item Response Theory. Dissertation, Jakarta: Jakarta State University.
Rustam, A (2019). Sensitivity and Accuracy of the Mantel-Haenszel Method and Standardization Method: Detection of Item Function of Item Functioning Differential. International Journal of Education and Literacy Studies. 7(3):28-37. https://doi.org/10.7575/aiac.ijels.v.7n.3p.28
Stoneberg, BD (2004). A Study of Gender Based and Ethnic Based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement Tests Applying the Simultaneous Bias Test (SIBTEST) and the Mantel Haenszel Chi Square Test. Internship in Measurement and Statistics. 1-15.
Swerdlik C (2010). Psychological Testing and Assessment: An Introduction to Test and Measurement. 7th ed. New York (USA): Mc Graw Hill.
Wahyuni, Ajeng (2022). Detection Of Gender Biased Using Dif (Differential Item Functioning) Analysis On Item Test Of School Examination Yogyakarta. JEP: Jurnal Evaluasi Pendidikan, Vol. 13(1), 46-49. https://doi.org/10.21009/jep.v13i1.26554
Zumbo, BD (2007). Three Generations of DIF Analysis: Considering Where It Has Been, Where It Is Now, and Where It Is Going. Language Assessment Quarterly. 4(2): 223-233. https://doi.org/10.1080/15434300701375832