Almufi Journal of Measurement, Assessment, and Evaluation Education (AJMAEE)
Website: http://almufi.com/index.php/AJMAEE
Email: almufi.ajmaee@gmail.com

LOGISTIC REGRESSION WITH ITEM RESPONSE THEORY (LRIRT): SENSITIVITY DETECTING DIFFERENTIAL ITEM FUNCTIONING

INFO PENULIS

Ahmad Rustam
Universitas Sulawesi Tenggara
ahmad.rustam1988@gmail.com
+6285399507330

INFO ARTIKEL

ISSN: 2798-0448
Vol. 1, No. 1, Juni 2021
http://almufi.com/index.php/AJMAEE

Saran Penilisan Referensi:

Rustam, A. (2021). Logistic Regression With Item Response Theory (LRIRT): Sensitivity Detecting Differential Item Functioning. *Almufi Journal of Measurement, Assessment, and Evaluation Education, 1* (1), 51-57.

Abstrak

Analisis butir sangat penting pada tes untuk mendapatkan informasi awal layak atau tidaknya suatu tes digunakan dalam penilaian. Salah satu kriteria baiknya suatu butir yaitu tidak terjadi diskriminasi atau menguntung pada golongan dalam menjawab benar suatu butir. Hal ini disebut dengan perbedaan fungsi butir disebut differential item functioning (DIF). Tujuan penelitian ini, apakah metode LRIRT lebih sensitif mendeteksi DIF dengan 2000 responden daripada 200 responden. Metode penelitian yang digunakan yaitu desain eksperimen, analisis yang digunakan two independent samples t-test. Data penelitian menggunakan hasil ujian nasional (UN) 2015. Hasil penelitian pertama,metode LRIRT lebih sensitif mendeteksi perbedaan fungsi butir (DIF) yang menggunakan 2000 responden daripada 200 responden. Kesimpulannya, metode LRIRT lebih sensitit deteksi DIF pada ukuran sampel 2000 daripada ukuran sampel 200.

Kata Kunci: Regresi logistik dengan IRT; differential item functioning; dif, bias

Abstract

Item analysis is very important in tests to obtain initial information whether or not a test is used in assessment. One of the good criteria for an item is that there is no discrimination or benefit to the group in answering correctly an item. This is called the difference in item function called differential item functioning (DIF). The purpose of this study is whether the LRIRT method is more sensitive in detecting DIF with 2000 respondents than 200 respondents. The research method used is experimental design, the analysis used is two independent samples t-test. The research data uses the results of the 2015 national exam (UN). The results of the first study, the LRIRT method is more sensitive in detecting differences in item function (DIF) which uses 2000 respondents rather than 200 respondents. In conclusion, the LRIRT method is more sensitive to DIF detection at a sample size of 2000 than a sample size of 200.

Key Words: Logistic Regression with IRT; differential item functioning; dif, bias

A. Introduction

Perbedaan Fungsi Butir (Differential Item Functioning)

Butir tes yang baik akan memberikan informasi atau hasil tes yang akurat. Ketika butir tidak berfungsi dengan baik, maka hasil yang akan digambarkan tentunya tidak baik. Salah satu faktor penyebab butir yang tidak baik yaitu terdapat ketidakseimbangan distribusi jawaban benar di antara kelompok peserta tes yang berbeda. Sehingga, hasil yang diperoleh tidak akurat dalam menggambarkan kemampuan siswa yang sebenarnya. Salah satu faktor yang mempengaruhi butir yaitu adanya bias pada butir atau yang lazim dikenal penelitian terbaru yaitu perbedaan fungsi butir atau differential item functioning (DIF). Zumbo (1999) menyatakan bahwa DIF yaitu istilah statistik yang digunakan untuk menggambarkan situasi di mana orang-orang dari satu kelompok menjawab suatu butir dengan benar lebih sering daripada orang yang sama-sama berpengetahuan dari kelompok lain. Deteksi DIF merupakan upaya untuk mengetahui apakah suatu butir tes bertindak adil atau tidak adil terhadap beberapa kelompok berbeda. (Camilli & Shepard, 1994) mencirikan perbedaan fungsi butir (DIF) sebagai jenis in-validitas yang merugikan satu kelompok lebih dari kelompok lain. Perbedaan fungsi butir (DIF) yang membedakan dua kelompok tersebut dikenal sebagai Differential item functioning (DIF).

Butir yang baik, mampu memberikan informasi yang akurat. Dalam hal ini bahwa butir tidak menguntungkan salah satu kelompok tertentu, sehingga butir tersebut akurat dalam pengambilan data kemampuan responden. (Hambleton, Swaminathan, & Rogers, 1991) menjelaskan bahwa perbedaan fungsi butir (DIF) yaitu sebuah butir dikatakan DIF, ketika beberapa individu dari kelompok berbeda yang memiliki kemampuan sama, akan tetapi tidak memiliki kemungkinan yang sama dalam menjawab butir dengan benar. Selain itu, dikemukakan bahwa differential item functioning (DIF) dapat juga diartikan sebagai perbedaan yang tidak diharapkan di antara beberapa kelompok ujian yang seharusnya hasil ujiannya sebanding berdasarkan atribut yang diukur oleh butir dalam tes yang dikerjakan (Wiberg, 2007). Differential item functioning (DIF) dapat diartikan bahwa di mana peserta ujian dari kelompok yang berbeda juga mempunyai kemungkinan-kemungkinan yang berbeda dalam menjawab suatu butir tes, setelah semua kemampuan dikontrol (Gierl, Khalid, & Boughton, 1999). Selain itu, DIF didefinisikan sebagai probabilitas yang berbeda dari peserta ujian dari kelompok yang berbeda namun dengan kemampuan yang sama merespons dengan benar terhadap butir (Ong, 2010). Prosedur deteksi DIF mengasumsikan sebuah perhitungan setelah disesuaikan dengan konstruk yang mendasari butir tersebut dimaksudkan untuk mengukur, dampak butir terjadi saat peserta ujian dari kelompok yang berbeda memiliki probabilitas yang berbeda untuk merespons dengan benar suatu butir (Camilli & Shepard, 1994). Terdapat dua bentuk perbedaan fungsi butir (DIF) yaitu DIF seragam (Uniform) dan DIF yang tidak seragam (Non-uniform).

DIF Seragam terjadi bila probabilitas untuk menjawab suatu butir dengan benar secara konsisten lebih tinggi untuk satu kelompok daripada kelompok lain pada semua tingkat kemampuan. Hal ini, ditandai dengan adanya dua butir Characteristic Curve (ICC) paralel. Dalam hal ini bahwa tidak ada interaksi antara tingkat kemampuan dan keanggotaan kelompok (Mellenbergh, 1982).

DIF tidak seragam terjadi bila perbedaan probabilitas untuk menjawab butir dengan benar berbeda pada arah yang berbeda untuk tingkat kemampuan yang berbeda untuk kelompok yang berbeda. Hal ini ditandai oleh dua ICC berpotongan. Dalam hal ini, ada interaksi antara tingkat kemampuan dan keanggotaan kelompok.

Faktor-faktor penyebab terjadinya bias butir soal dalam pelaksanaan tes menurut Crocker & Algina, (1986) bahwa ketika membahas tentang pengelompokan minoritas dan mayoritas dalam pelaksanaan tes, menyatakan bahwa kedua istilah tersebut di atas dapat didefinisikan dengan adanya dua kelompok yang disebabkan oleh perbedaan karakteristik yang ditinjau dari segi ras, latar belakang budaya, umur, dan cacat secara fisik. Hal ini ditegaskan pula (Jensen, 1980) bahwa muncul bias pada butir benar-benar terjadi disebabkan oleh faktor ras dan sex.

Nilai yang diperoleh dari hasil tes dipaparkan dengan tujuan bahwa untuk memberikan informasi tentang besaran atau dimensi yang diukur oleh tes. Kadang kala skor pada hasil tes tidak memberikan informasi yang akurat mengenai peserta tes. Paket tes dapat menjadi tidak berfungsi dengan baik kualitasnya disebabkan adanya perbedaan fungsi pada butir tes.

Berdasarkan uraian tersebut, kemampuan metode deteksi untuk mengecek ada tidaknya perbedaan fungsi butir (DIF) pada tes setiap butir, sangat diharapkan untuk melakukan proses pengukuran untuk dilakukan agar ketidakadilan atau kerugian pada kelompok tertentu dapat dihindarkan serta kemampuan siswa terukur secara objektif.

Tujuan penelitian ini,untuk menguji metode LRIRT, apakah lebih sensitif mendeteksi DIF dengan 2000 responden daripada 200 responden.

Regresi Logistik dengan IRT

Swaminathan & Rogers (1990) menuliskan bentuk alternatif formula yang ekuivalen dengan metode logistik regression yang digunakan untuk deteksi DIF. Kemampuan peserta yang digunakan dalam analisis yaitu kemampuan peserta yang di estimasi dengan model logistik tiga parameter (L3P). Hal ini, sejalan dengan hasil penelitian Crane, Belle, & Larson (2004) yang mengganti jumlah skor kemampuan peserta dengan kemampuan peserta atau theta (θ) hasil estimasi menggunakan L3P.

Metode ini dapat mendeteksi uniform DIF dan nonuniform DIF sekaligus. Model probabilitas sebagai berikut,

(1) $$P(u=1)=\frac{e^{z}}{\left(1+e^{z}\right)}$$

dimana, P(u=1) adalah peluang responden menjawab benar suatu butir tes,

(2) $$z=\tau_0+\tau_1\theta+\tau_2 g+\tau_3(\theta g)$$

dimana, θ adalah tingkat kemampuan peserta tes hasil estimasi model logistis tiga parameter (L3P), g adalah kelompok peserta yang diberi kode 1 (untuk kelompok fokal) dan 2 (untuk kelompok referensi), sedangkan θg adalah perkalian dari dua variabel independen yaitu θ dan g. Nilai signifikansi τ_2 menunjukkan adanya perbedaan kelompok dalam kinerja pada butir atau disebut dengan DIF Uniform, dan nilai signifikansi τ_3 menunjukkan adanya interaksi antara anggota kelompok dan kemampuan atau disebut dengan DIF nonuniform.

Uji signifikansi DIF menggunakan uji Wald dengan bantuan program SPSS for windows (Field, 2000) Sebagai berikut:

(3) $$Wald=\frac{b}{SE_b}$$

dimana, b adalah estimated regression coefficients and SE adalah standard errors. Uji ini menggunakan chi-square distribution.

Sensitivitas

Sensitivitas pertama kali diperkenalkan Yerushalmy pada pengukuran kesehatan bahwa Sensitivitas yaitu kemampuan mendiagnosis secara benar pada orang yang sakit, berarti hasil tesnya positif dan memang benar sakit (Yerushalmy, 1947). Hal ini, dikaitkan dengan pengukuran bahwa sensitivitas merupakan proporsi butir yang positif DIF dalam populasi dan setelah diidentifikasi oleh metode deteksi ternyata benar butir tersebut DIF. Dengan kata lain bahwa sensitivitas merupakan kemungkinan butir DIF terdeteksi dengan benar atau probabilitas setiap butir yang DIF teridentifikasi benar dengan metode deteksi DIF. Dalam dunia pengukuran atau psikometri, konsep tersebut digunakan untuk mendeteksi butir yang

terindikasi memiliki DIF dan dikenal dengan true positive. True positive yaitu apabila butir yang benar DIF dan setelah dideteksi dengan metode tertentu dan hasilnya positif DIF. True positive sering disebut dengan kekuatan (power) suatu metode deteksi DIF. Selain itu, dikenal juga tingkat kesalahan tipe II (type II error rate). Tingkat kesalahan tipe II (type II error rate) adalah metode deteksi pada butir yang benar terdapat DIF dan hasil deteksi butir tersebut tidak terdapat DIF.

Loong (2003) menerangkan tentang sensitivitas bahwa sensitivitas = TP / TP + FN, di mana TP adalah jumlah positif yang sesungguhnya dan FN adalah jumlah negatif palsu. Berdasarkan tabel tersebut, sehingga formulasi sensitivitas yang dimaksud dalam penelitian ini yaitu

$$Sensitivity = \frac{\sum TruePositive}{\sum TruePositif + \sum FalseNegative} \quad (4)$$

Berdasarkan formula tersebut, suatu metode deteksi DIF semakin sedikit negatif salah yang terdeteksi, maka semakin sensitif metode tersebut. Sebaliknya, suatu metode deteksi DIF semakin banyak negative salah yang terdeteksi, maka semakin kurang sensitif metode tersebut.

Sensitivitas yang rendah diakibatkan dari metode deteksi DIF yang melewatkan banyak butir yang mengandung DIF. Hal ini, dapat dikatakan bahwa suatu metode deteksi DIF dengan sensitivitas yang rendah akan meningkatkan beberapa jumlah negatif salah atau false negative (FN).

Ukuran Sample

Ukuran sampel menjadi salah satu pertimbangan penting dalam menerapkan model untuk mengestimasi parameter khususnya pada Item Response Theory (IRT). Estimasi akan baik ketika sampel yang digunakan dapat memenuhi syarat untuk model yang akan diterapkan.

Ukuran sampel minimal yang wajar untuk model 1 parameter logistik (L1P) sebesar 200 responden (Linacre, 2005). Dengan demikian, minimum 200 responden dalam Item Response Theory (IRT) sudah cukup akurat. Sebagaimana juga ditegaskan oleh (Crocker & Algina, 1986) bahwa untuk menganalisis butir menggunakan Item Response Theory (IRT) sebaiknya minimal 200 responden.

Estimasi yang baik pada dasarnya sangat dipengaruhi oleh ukuran sampel. Ukuran sampel /responden yang semakin besar/banyak, maka akan semakin stabil hasil estimasi parameter butir. Sebaliknya, jika semakin kecil ukuran sampel /responden, maka kestabilan estimasi akan menurun. Seperti yang dikemukakan (Hulin, Lissak, & Drasgow, 1982) bahwa estimasi parameter kemampuan dan butir kurang akurat pada ukuran sampel kecil khususnya ketika respons butir dibangkitkan oleh model logistik tiga parameter (L3P). Selain itu, (He & Wheadon, 2008) melaporkan penelitiannya bahwa pengaruh ukuran sampel terhadap tingkat stabilitas dan akurasi parameter model bahwa dipengaruhi oleh ukuran sampel, jumlah kategori butir dan distribusi skor kategori dalam butir. Secara umum bahwa sampel yang semakin mendekati ukuran populasi, maka akan semakin baik hasilnya.

Pada model IRT sebagaimana yang dijelaskan sebelumnya bahwa membutuhkan ukuran sampel setidaknya 200 responden. Akan tetapi, untuk mendapatkan kestabilan yang baik maka perlu untuk menambah ukuran sampel. Estimasi dengan error yang semakin kecil juga dipengaruhi jumlah sampel. Sebagaimana hasil penelitian (Reise & Yu, 1990) memaparkan bahwa sekitar 500 peserta ujian dibutuhkan untuk mencapai root mean square error (RMSE) di bawah 0,10 dan ukuran sampel sebesar 2.000 akan mencapai nilai RMSE sekitar 0,05. Selain itu, (Reise & Yu, 1990) menjelaskan bahwa ukuran sampel memiliki pengaruh yang paling banyak diputuskan pada korelasi, menggambarkan bahwa korelasi r = 0,77 dicapai pada ukuran 250 peserta ujian, dan nilai korelasi meningkat sampai 0,95 untuk ukuran 2.000 peserta ujian. Untuk itu, sekitar 1.000 peserta ujian dibutuhkan untuk mempertahankan rata-rata estimasi sebenarnya dari korelasi 0,90. Menurut (Hulin et al., 1982) merekomendasikan agar untuk 500 responden untuk dua parameter model (L2P) menggunakan 500 responden, dan 1000 responden untuk tiga parameter model (L3P).

B. Methodology

Metode penelitian yang digunakan yaitu desain eksperimen dengan rancangan perlakuan. Adapun variabel penelitian terdiri dari variabel bebas dan variabel terikat. Adapun variabel terikat dalam penelitian ini yaitu banyaknya butir positif benar atau true positive (TP) atau sebut dengan power. Sedangkan, variabel bebas yaitu metode deteksi perbedaan fungsi butir (DIF) yang terdiri dari metode logistic regression with item response theory (LRIRT).

1. Data

Data utama yang digunakan merupakan data respons siswa yang berbentuk skor, dimana bentuk skor hasil siswa berupa respons "0" dan "1" dengan panjang tes atau jumlah butir dalam tes sebanyak 40 butir. Sumber data dari Pusat Penilaian dan Pengujian Pendidikan Departemen Pendidikan Nasional (PUSPENDIK) yang berupa data respons hasil ujian nasional (UN) siswa pada tahun 2015.

2. *Population and Sample*

Populasi dalam penelitian ini yaitu butir tes bentuk pilihan ganda dan peserta tes. Yang menjadi populasi peserta tes dalam penelitian yaitu para peserta tes ujian nasional (UN) kabupaten Bone dan Kabupaten Luwuk Timur Kabupaten Bunggai sebanyak 3.054 peserta tes.

Dari populasi tersebut diambil sampel/peserta tes untuk variabel pertama yaitu jenis kelamin ditentukan laki-laki sebagai kelompok referensi diambil sebanyak (1000 dan 100) respons dan sampel perempuan sebagai kelompok fokus sebanyak (1000 dan 100) respons.

Jumlah replikasi yang telah dilakukan dalam penelitian ini sebanyak 250 kali secara acak. Sedangkan populasi butir tes yaitu butir soal yang digunakan dalam ujian nasional (UN) sebanyak 40 butir soal. Sampel butir yang digunakan yaitu sebanyak 25 butir yang diambil secara acak dan cocok dengan model setelah dilakukan replikasi.

3. *Prosedur Penelitian*

Prosedur penelitian ini dimulai dengan mempersiapkan data utama yaitu data berupa skor pekerjaan siswa SMP pada ujian nasional tahun 2015 yang diperoleh dari Pusat Penilaian dan Pengujian Pendidikan Departemen Pendidikan Nasional (PUSPENDIK). Data berupa skor yang berbentuk nol (0) dan satu (1) dengan panjang perangkat tes yaitu 40 butir. Dari 40 butir tes dideteksi DIF menggunakan bantuan program BILOG MG. Tujuan deteksi DIF ini untuk mendapatkan informasi awal, butir mana saja yang terdeteksi DIF atau tidak DIF. Jumlah Respon yang digunakan untuk deteksi DIF masing-masing kelompok referensi dan kelompok fokus adalah 1527 respons, jadi total peserta tes sebanyak 3054.

Jumlah respons peserta tes masing-masing kelompok referensi dan fokus adalah 1527. Kemudian, diambil secara acak respons sebanyak 1000 dan 100 respons sebanyak 75 replikasi untuk masing-masing kelompok referensi dan kelompok fokus menggunakan bantuan SPSS. Kemudian, setelah selesai pengambilan respon, selanjutnya menggabungkan data referensi dan fokus menjadi 2000 dan 200 respon, Setelah digabung kemudian mengestimasi parameter kemampuan dan parameter butir. Jumlah respon yang digunakan untuk estimasi butir dan estimasi kemampuan adalah 2000 dan 200 respons.

Estimasi dilakukan dengan menggunakan program BILOG. Dari hasil estimasi butir yang cocok model hanya 25 dari 40 butir. Total replikasi 75, namun hanya 15 replikasi yang memuat butir yang cocok model pada nomor yang sama yaitu sebanyak 25 butir.

Dari 25 butir dilakukan estimasi parameter untuk masing-masing kelompok referensi dan fokus. Setelah itu, mendeteksi butir yang memuat DIF dengan menggunakan dua metode deteksi yaitu metode LRIRT. Deteksi DIF dengan metode LRIRT menggunakan bantuan program SPSS.

4. Teknik Analisis Data

Berdasarkan hasil analisis uji normalitas menggunakan Lilliefors pada data true positive (TP) untuk setiap kelompok diperoleh tidak signifikan. Sehingga, dalam analisis hipotesis menggunakan statistik non parametrik yaitu uji beda rata-rata Mann- Whitney Test. Uji perbedaan rata-rata kelompok ditinjau dari data nilai true positive (TP) atau positif benar. Uji beda rata-rata yang digunakan yaitu Mann-Whitney Test.

C. Finding and Discussion

Hasil pengujian hipotesis berdasarkan analisis statistik inferensial menggunakan uji perbedaan dua sampel independen untuk masing-masing hipotesis akan di tampilkan sebagai berikut.

Berdasarkan hasil analisis statistik bahwa hasil analisis statistik uji perbedaan rata-rata dua kelompok sampel, diperoleh hasil perhitungan statistik bahwa nilai Mann-Whitney U

sebesar 0.000 dan nilai probabilitas (sig.) lebih kecil dari α=0.05 yaitu 0.000, berarti hipotesis H0 ditolak. Dengan demikian, metode LRIRT lebih sensitif mendeteksi perbedaan fungsi butir (DIF) yang menggunakan 2000 responden daripada 200 responden.

Sensitivitas diartikan sebagai proporsi butir yang positif DIF dalam populasi dan setelah diidentifikasi oleh tes atau metode deteksi DIF ternyata benar butir tersebut DIF. Dengan kata lain bahwa sensitivitas merupakan kemungkinan butir DIF terdeteksi dengan benar atau probabilitas setiap butir yang DIF teridentifikasi benar dengan metode deteksi DIF. Hasil penelitian ini menguji hipotesis yang berkaitan dengan metode deteksi DIF dan dua ukuran sampel yaitu 2000 dan 200. Hasil analisis hipotesis menunjukkan bahwa metode LRIRT lebih sensitif mendeteksi perbedaan fungsi butir (DIF) yang menggunakan 2000 responden daripada 200 responden.

Hasil analisis yang menunjukkan bahwa metode LRIRT lebih sensitif mendeteksi perbedaan fungsi butir (DIF) yang menggunakan 2000 responden daripada 200 responden. Ditinjau dari ukuran sampel, hasil analisis true positive (TP) atau disebut dengan kekuatan (power) metode deteksi DIF, semakin meningkat dengan bertambahnya ukuran sampel. LRIRT merupakan metode regresi logistik yang menggunakan nilai kemampuan siswa atau theta (θ) hasil estimasi dari IRT, hal ini sangat berkaitan dengan ukuran sampel. Sehingga ketika ukuran sampel semakin besar, maka hasilnya cenderung lebih baik dan sensitivitas dari regresi logistik akan bertambah.

Sensitivitas deteksi DIF pada metode LRIRT untuk ukuran sampel 2000 hasilnya mencapai peningkatan dua kali lebih besar dari ukuran sampel 200. Nilai rata-rata sensitivitas LRIRT pada ukuran sampel 2000 yaitu sebesar 0.903, sedangkan rata-rata sensitivitas LRIRT pada ukuran sampel 200 yaitu sebesar 0.376. Hal ini, sejalan dengan hasil penelitian (Reise & Yu, 1990) bahwa bertambahnya ukuran sampel akan memperkecil nilai root mean square error (RMSE), untuk ukuran sampel 500 dapat mencapai nilai RMSE yang kecil di bawah 0,10 dan menegaskan pada ukuran sampel sebesar 2.000 dapat mencapai nilai RMSE sekitar 0,05.

D. Conclusion

Kesimpulan, metode LRIRT lebih sensitit deteksi DIF pada ukuran sampel 2000 daripada ukuran sampel 200.

E. References

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. SAGE Publications Inc.

Crane, P. K., Belle, G. Van, & Larson, E. B. (2004). Test bias in a cognitive test : differential item functioning in the CASI. *Statistics in Medicine*, *23*, 241–256. https://doi.org/10.1002/sim.1713

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: CBS college publishing.

Engelhard, G. (2009). Using Item Response Theory and Model–Data Fit to Conceptualize Differential Item and Person Functioning for Students With Disabilities. *Educational and Psychological Measurement Volume*, *69*(4), 585–602. https://doi.org/10.1177/0013164408323240

Field, A. (2000). *Discovering Statistics Using SPSS*. London: SAGE Publications Inc.

Gierl, M., Khalid, S. N., & Boughton, K. (1999). Gender Differential Item Functioning in Mathematics and Science : Prevalence and Policy Implications. In *Improving Large-Scale Assessment in Education* (pp. 1–25). Canada: Centre for Research in Applied Measurement and Evaluation University of Alberta Pap.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. California: SAGE Publications Inc.

He, Q., & Wheadon, C. (2008). The Effect of Sample Size on Item Parameter Estimation For The Partial Credit model. *Centre for Education Research and Policy*, (3644723).

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of Two- and Three-Parameter Logistic Item Characteristic Curves : A Monte Carlo Study. *Applied Psychological Measurement*, *6*(3), 249–260.

Ironson, G. H. (1983). *Using Item Response Theory to Measure Bias", in Application of Item Response Theory, ed. Ronald K. Hambleton*. Vancouver: Educational Research Institute of Bristish Columbia.

Jensen, A. R. (1980). *Bias in Mental Testing*. New York: A Division of Macmillan Publishing Co., Inc.

Linacre, J. M. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. *Rasch Measurement Transactions*, *19*(3), 1032.

Loong, T. (2003). Understanding sensitivity and specificity with the right side of the brain. *BMJ*, *327*, 716–719.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias, *7*(2), 105–118.

Ong, Y. M. (2010). *Understanding Differential Functioning By Gender in Mathematics Assessment*. University of Manchester for the degree of Doctor of Philosophy.

Reise, S. P., & Yu, J. (1990). Parameter Recovery in the Graded Response Model Using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133–144.

Wiberg, M. (2007). Measuring and Detecting Differential Item Functioning in Criterion-Referenced Licensing Test: A Theoretic Comparison of Methods. *EM*, (60).

Wilson, M. (2005). *Constructing Measures : An Item Response Modeling Approach*. London: Lawrence Erlbaum Associates. https://doi.org/10.1080/15305050802007117

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Yerushalmy, J. (1947). Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques. *Public Health Reports*, *62*(40), 1432–1449.

Zumbo, B. D. (1999). *A handbook of 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.