Dirasat: *Educational Sciences*, Volume 48, No. 3, 2021

Deanship of Scientific Research

The Implication of Wald Test Anchor-All-Test-All Procedure for Anchor Selection

Yahia Alsmadi

School of Educational Sciences, The University of Jordan, Jordan.

Received: 29/7/2020
Revised: 14/8/2020
Accepted: 22/9/2020
Published: 1/9/2021

Citation: Alsmadi, Y. (2021). The Implication of Wald Test Anchor-All-Test-All Procedure for Anchor Selection. *Dirasat: Educational Sciences*, *48*(3), 428-436. Retrieved from https://dsr.ju.edu.jo/djournals/index.php/Edu/article/view/2885

Abstract

This study aims to provide information on whether the more practical and time saver anchor-all-test-all procedure (AATA) strategy for anchor selection can demonstrate similar performance or outperform the well known all-others-as-anchors (AOAA) procedure under certain conditions.

FlexMIRT-3 and IRTLRDIF-2 was utilized for Wald's x 2 AATA and IRT-LR AOAA. All the parameters were constrained between the groups equally for the estimation of focal group distribution. These parameters were estimated through a full model that confined the focal group mean and SD values from the baseline model. The X2 statistics have been used to evaluate the differences between two sets of item parameters of two different groups.

DIF tests results were dependent on the anchor selection strategies employed. The results have revealed that appropriate anchor-selection not only depends on the sample size, but it has an association with the proportion of DIF items and the direction of DIF along with the length of the anchor. Test depicts the efficacy of the Wald method for the DIF.

This study suggests future researchers compare the different modeling approaches for the detection of DIF methods such as multiple indicators multiple causes (MIMIC) modeling following variant simulation techniques. In addition, future studies are recommended to analyze the efficiency of the anchor selection strategies concerning different groups and times.

Keywords: Wald X2 Test, DIF testing, anchor items, anchor-all-test-all, test validity.

تطبيق اختبارواالد كطريقة فاعلة في اختيار الفقرات الر ابطة

يحيى الصمادي

الجامعة الأردنية ، الأردن.

ملخّص

الأهداف: هدفت هذه الدراسة إلى توفير معلومات حول ما إذا كانت استراتيجية إجراء اختبار الكل (AATA) الأكثر عملية وتوفيراً للوقت لاختيار الفقرات الرابطة يمكن أن يظهر أداءً مشابهًا أو يتفوق في الأداء على إجراء جميع الفقرات الباقية كفقرات رابطة (AQAA) في ظل ظروف معينة.

المنهجية: تم استخدام FlexMIRT-3 و IRTLRDIF-2 مع Wald's x 2 AATA و IRT-LR AOAA. تم تقييد جميع المعالم بين المجموعات بالتساوي لتقدير توزيع المجموعة المرجعية. تم تقدير هذه المعالم من خلال نموذج كامل حصر متوسط المجموعة المستهدفة وقيم الانحراف المعياري من النموذج الأساسي. تم استخدام إحصائيات X2 لتقييم الاختلافات بين مجموعتين من معالم الفقرات لمجموعتين مختلفتين.

النتائج: أظهرت النتائج أن اختبارات الأداء التفاضلي DIF تعتمد على استراتيجيات الاختيار المستخدمة، وكشفت أن اختيار الفقرات الرابطة المناسب لا يعتمد فقط على حجم العينة، بل يرتبط أيضًا بنسبة الفقرات ذات الأداء التفاضلي واتجاهه، وأيّدت النتائج استخدام اختبار والد WALD كطريقة فاعلة في اختيار الفقرات الرابطة.

الخلاصة: تقترح هذه الدراسة أن يقارن الباحثون المستقبليون طرق النمذجة المختلفة للكشف عن طرق DIF مثل المؤشرات المتعددة ، ونمذجة الأسباب المتعددة (MIMIC) باتباع تقنيات المحاكاة المتغيرة. أيضًا، ويوصى بتحليل فاعلية إجراءات اختيار الفقرا الرابطة المتعلقة بالمجموعات والأوقات المختلفة.

الكلمات الدالة : اضطراب القلق، الرضا عن الحياة، الطلبة المنذرين بالفصل في جامعة مؤتة

Introduction

Differential Item Functioning (DIF) is the invariance assumption violation in the Item Response Theory (IRT) models (Woods, Cai, & Wang, 2013). It occurs when the positive response probability for the examinees at the same ability level differs in different groups (Battauz, 2019). DIF also occurs in the misspecification of the latent ability space. For assessing the group's ability difference, invariant items parameters are needed objectively. In case violence of the invariance assumption occurs, the subgroups are accounted for different item characteristic curves (Kopf, Zeileis, & Strobl, 2015).

For example, when the multidimensional ability of the groups differs, the items used for differentiating among the abilities following a unidimensional score, might flag items as DIF's (Bastug, 2016). Previous studies have proposed various methods for DIF detection, including Mantel–Haenszel test, item response theory, likelihood ratio test, improved version of Lord's Wald X^2 Test, and logistic regression) (Magis, Béland, Tuerlinckx, & De Boeck 2010).

Lord's chi-square test is the most common procedure, which shows the estimation advantages concerning the item parameters for each group once, given that the anchor item selection occurs in the following step (second step). This test was originally developed for the DIF detection between two groups, which was then expanded to multiple groups (Kim, Cohen, & Kim, 1994). To match the subjects on the latent factor, these methods need a group-invariant anchor subset before the DIF analysis is carried out. Inaccuracy and Type, I and II error, occurs if the group-invariant anchor subset is polluted (Woods, Cai, & Wang, 2013).

Therefore, DIF analysis is a critical step because it may result in the inaccuracy of parameters' estimation for identifying anchor set/item. In the past, various empirical anchor-selection methods have evolved and investigated. The procedure of IRT-LR (item response theory-likelihood ratio) all-others-as-anchors (AOAA) yields less power and a greater chance of type I error to occur (Wangs & Woods, 2017). Therefore, this study has investigated whether the procedure of IRT-LR AOAA could be replaced by the Wald AATA (anchor-all-test-all) procedure, as this procedure saves time and needs just two model-fitting under various conditions. Moreover, this procedure has been tested for the generation of high power and well-controlled type 1 error during the DIF analysis.

The study has evaluated the association of type I error and statistical power with DIF analysis by utilizing the anchors that are previously selected by the corresponding strategy. Moreover, this study has also investigated the association between the variables that can be manipulated (like sample size and length of the test) and the results of DIF analysis. Thus, the study has focused on the alternative strategies that are utilized for anchor selection that can overcome the limitations associated with the IRT-LR AOAA procedure.

Previously, various simulation techniques are conducted concerning the technical DIP issues, and where the methodological problems continue to emerge (Zumbo, 2007). Various studies have evaluated the DIF method accuracy using item response theory (Cao, Tay, & Liu, 2017; Chun, Stark, Kim, & Chernyshenko, 2016; Hou, la Torre, & Nandakumar, 2014; Woods, Cai, & Wang, 2013; Wang & Shih, 2010; Wang, Shih, & Yang, 2009; Woods, 2009); however, these appear to be limited concerning the DIF detection comparison, which makes their capabilities questioning. Studies have indicated efficacy of these methods, where IRT DIF detection method continues to be most capable; however, these require continuous comparison in simulation research (Woods, Cai, & Wang, 2013; Tay et al., 2015; Cao, Tay, & Liu, 2017; Chun et al., 2016).

This is the reason that present study conducts a comparison of the MIMIC approach for correct identification of DIF uniform and nonuniform with the IRT based likelihood ratio tests (IRTLR) and Wald test approaches (Tay, Meade, & Cao, 2015; Woods et al., 2013; Cao et al., 2017). Recent research has revealed that DIF detection method is the most capable, though its use in the comparison for the simulation research has remained limited (Tay et al., 2015; Chun et al., 2016; Cao et al., 2017), which is required to be verified through the conditional testing. This comparison is likely to aid the practitioners and the researchers for determining an optimized method. The study, therefore, analyzes alternative strategies for anchor selection that help overcome the limitations of IRT-LR.

Moreover, various studies have shown an inclination to the testing of the DIF across several groups at the same time for the identification of the assessment bias. This method also serves as an alternative to the conventional IRT procedure as

it is easy for the accommodation of the background variables as well as their interaction in the absence of the large samples (Kim, Yoon, & Lee, 2012). The specification of the anchor item is also dictated by the Wald-2 test. In case, the studied items function differently, the empirical method for the selection of anchor is generally part of the DIF literature, which assists in the correct determination of the Type I errors.

The Problem of the Study

The problem of this study could be stated as to provide information on whether the more practical and time saver AATA strategy for anchor selection can demonstrate similar performance or outperform the well known AOAA strategy under certain conditions.

Previous Studies and Literature Review

The selection and identification of an accurate group-invariant anchor subset play an important part in the process of DIF analysis. In the following sections, the relevant previous studies and literature well be reviewed and demonstrated within each section.

Procedures for Anchor Selection

The procedure of IRT-LR AOAA is known as the most popular procedure for anchor selection (Woods, 2009; Thissen, Steinberg, & Wainer, 1993). With the equalization of the corresponding reference and focal parameters of the group, the procedure of IRT-LR AOAA fits a baseline model to the data. The procedure can be paired with the MingG^2 criterion or NonsigMaxA criterion (Woods, 2009). The change in parameters is associated with the analyzed items; whereas, the multiple augmented models are shaped according to the data of other items such as anchors. The IRT-LR test is specific as it can limit the known and verified anchors among different groups Wang & Woods, 2017).

The IRT-LR AOAA method has acted as a baseline model with the corresponding reference and parameters of the focal group equally. All the selected items are treated as anchors to fit the multiple augmented models; whereas, different parameters of analyzed items vary from time to time. A set of known anchors is constrained in the baseline between the groups during the IRT-LR testing (Wang & Woods, 2017). To investigate the equal constraints on each item, the nested models are compared. The retaining of G^2 statistics as anchors is associated with the G^2 non-significance criterion (Thissen, Steinberg, & Wainer, 1993). The IRT-LR AOAA can be easily paired with MinG^2 criterion or Nonsig criterion.

MinG^2/Minx^2 Criterion

Based on the G^2 values, the MinG^2/Minx^2 criterion picks up anchors by classifying them in ascending order (smallest to largest). However, the items that possess the smallest G^2 values are selected as anchors. The anchor selection is based on various factors that include the sample size and test length. It has been suggested that it is not appropriate to select more than 25% of the total items as anchors (Meade & Wright, 2012). The MinG^2 approach has reflected the increased DIF effects associated with an increase in the G^2 values. The MinG^2 approach possesses a distinctive property as it does not consider any non-significant G^2 statistics before ranking the anchors (Wang & Woods, 2017). The origination of MinG^2 approach is associated with the increased G^2 values that are reflected through the increased DIF effects. This criterion is distinct as it does not need significance of G^2 statistic before ranking. Therefore, MinG^2 (Wald version) is also known as Minx^2 criterion.

NonsigMaxA Criterion

The NonsigMaxA Criterion is responsible for the ranking of non-significant G^2 values based on reference group estimation that discriminates the parameters in descending order (Lopez Rivas, Stark, & Chernyshenko, 2009). The items with increased discriminatory parameters are determined as anchors. Therefore, this specific pattern is effective in accurate determination of anchor items that possess increased discrimination. In a situation where no similarity between the latent

distributions is detected, the anchor items with increased discriminations are least prone to Type I and Type II errors during the IRT-LR testing (Ankenmann, Witt, & Dunbar, 1999).

The highly discriminated items are not usually identified as non-significant when an increased probability of Type I error exists. These items are retained as non-significant when the DIF effect is exhibited. A comparison of anchor-selection methods performance and their variants revealed that the application of NonsigMaxA criterion yielded better results as compared to the Nonsig criterion (Meade & Wright, 2012). It has been suggested to couple the IRT-LR AOAA procedure with an alternative criterion to maintain high power and minimize type I error during the DIF analysis.

Procedure of IRT-LR AOAA and its Concerns

The procedure of anchor selection is dependent on the implementation of IRT-LR AOAA techniques. Although the technique is much popular, it owns certain drawbacks. The procedure of IRT-LR AOAA needs multiple model fittings. Therefore, it consumes much time with larger sample size and lengthy tests (Thissen, 2001). The procedure of IRT-LR AOAA is not easy to compare among three or more groups because a huge amount of model fittings is needed for the comparison of groups in pairs. The addition of DIF items resulted in the decreased specification of the augmented models that are used to carry out the IRT-LR AOAA procedure. The inclusion of DIF items decreases the specification of augmented models, utilized during the IRT-LR AOAA procedure (Maydeu-Olivares & Cai, 2006).

Improved Version of Lord's Wald X^2 Test

The Lord's Wald X2 Test just needs a single model fitting during the DIF analysis. The parameters that are needed to be studied can be freely estimated for each group based on the latent scores using this test, as it utilizes the anchors for joining the groups on a single latent scale (Woods, Cai, & Wang, 2013). The Wald test is used to investigate the groups' connection situated on a single latent scale; whereas, the estimation of studied parameters is performed for each group according to their latent scores. There is no chance of any theoretical problem to arise during the specification of augmented models (the model including all the items other than the one analyzed) items except the item being analyzed) because the Wald test is not associated with the comparison between nested models (the model that use the similar variables for another model, however, specifies the additional parameter to be estimated), for their G2 test statistics. Therefore, the Wald test can conduct the DIF analysis of various groups by introducing a different contrast of coefficient matrix (Langer, 2008; Woods, Cai, & Wang, 2013). The improved version of the Wald test model has increased the expected margin of the procedure to estimate various parameters. Moreover, it has also improved the concurrent calibration approach and supplemented the expectation-maximization (SEM) algorithm for the calculation of the error covariance matrix (Kolen & Brennan, 2004).

Study Methods and Procedures

Data Generation

Quantitative techniques have been opted, including the AOAA-MinG2 strategy, IRT-LR test, and Logistic Regression Models. Alternative anchor-selection strategies have been retrieved through regression models. Despite the commonly used traditional factorial analysis methods for data generation and models specification, the study has implicated a randomized simulation mechanism, in which the factor levels were drawn randomly. The evaluation criteria integrate the type I and error as well as power. Type 1 error computation include identification of non-DIF item as DIF following its division by the non-DIF scale items.

Study Simulation Design

The application of advanced and improved speed for computation, the simulation design will be applied for summarizing results according to the advanced modeling techniques. This will lead to a broader approach regarding the standard practice of artificially categorized factors with a high risk of losing information from a continuous scale. The

artificial categorization of continuous variables is used for saving simulation time during the generation of data for a specified number of levels. The study has utilized advanced computing speed for conducting simulation procedure and summarizing the results. Moreover, different factors have been randomly drawn from the previously specified values. This type of design facilitates results modeling to yield the best outcomes under optimal conditions. Therefore, the study conducted 10,000 replications to obtain diversified simulated values and improved results.

Fixed Factors. In this study, the 2PL model was used for two equal sample-sized groups. Reference latent trait levels, and focal groups were spread as Θ_R~ (0, 1) and Θ_F~N (- 0.6, 1). The latent mean for the reference group was higher than the focal group for testing the robustness of the anchor selection strategies and impact in comparison to DIF effects.

Varying Factors. From the discrete uniform distribution, the sample size was randomly drawn from 400 to 4000, where the sample size of each group (nSubjects), ranging from 250 to 2500. For each replication, the test length value was randomly drawn from the distribution of discrete uniform ranging from 5 to 50. For replication, the differential functioning properties were present in at least one item where the DIF items maximum number did not go more than 80 percent of nItems, where every replication was guaranteed to have group-invariant items of 20 percent or more. The limitations were determined to depict a different range of possible conditions. For item parameters, the difficulty parameter for the reference group was derived from b_R~N (0.1, 1.3^2) having truncations at 2.3 and -2.4, while these truncations were at 2.0 and 0.4 for the discriminatory parameters drawn from a_R~N (0.8, 0.4^2). The items parameter distributions, as well as truncations for the decisions, were informed for 589 dichotomous items derived from the educational and psychological empirical studies (Childs, Dahlstrom, & Panter, 2000; Lord, 1968; Marsman, Waldorp, & Maris, 2017). For the DIF effect magnitudes, uniform distributions were used for drawing the item parameters differences $|b_{DIF}| \in |0.2, 0.6|$ and $|a_{DIF}| \in |0.6, 0.6|$. Theoretical findings show that uniform DIF exists for b_{DIF}, while nonuniform for nonnegligible a_{DIF}. The parameter differences were added for obtaining the focal groups' item parameters (such as DIF effect magnitudes) for the corresponding reference group items. The DIF magnitude combination, there exists a difference of at least 0.2 units in the location parameters, whereas the group's subset items also differed based on discrimination parameters.

Study Procedures

FlexMIRT version 3 and IRTLRDIF version 2 will be utilized for Wald's x 2 AATA and IRT-LR AOAA, respectively, as software applications. Firstly, the data was analyzed using Wald's X^2 AATA by fitting to the baseline model. All the parameters were constrained between the groups equally for the estimation of focal group distribution. Moreover, these parameters were estimated through a full model that confined the focal group mean and SD values from the baseline model. The X^2 statistics have been used to evaluate the differences between two sets of item parameters of two different groups. The anchor set, selected through anchor-selection strategy, was used to carry out the four follow-up DIF tests during each replication.

Limitations of the Study

The generalizability of this study's results and conclusions are limited to the procedures, methods, strategies and criteria chosen and depicted in this study. For example, the study implemented the 2PL model to generate dichotomous data, and therefore further studies are needed to replicate its results using multiple parameters models and multiple grade data as well as different procedures and methods of anchor selection.

Study Results and Discussion

Underbalanced conditions, the anchor selection procedure of AOAA is more accurate as compared to the AATA procedure. The main concern is to evaluate which procedure performs well in terms of achieving maximum power with a

well-controlled chance of type I error. The results have been analyzed and compared with other studies. When a single item is designated as an anchor, the IRT-LR generates extremely low power (Meade & Wright, 2012).

Known Focal Group

The item parameters based on fixed focal group mean and standard deviation were estimated through Wald's X^2 AATA, and the biasness in estimating the distribution of the focal group has been recorded. The average mean for the focal group was estimated to be – 0.36 (ranging between 1.03 – 0.44), and the average SD for focal group was estimated to be 0.81 (ranging between 0.58 – 1.13). The mean of squared errors for the focal group was calculated to be 0.26, and SD was 0.16

. Biased results have been obtained after estimating the focal group distribution because the procedure for all the factors has been used as anchors for the estimation of focal group distribution.

Estimation of Anchor Selection Accuracy

The accuracy of anchors was confirmed through the Nonsig criterion associated with AOAA and AATA procedure to figure out the potential stimulation errors. Low accuracy (31.26% with a significance value of 0.05) was estimated, as AATA was associated with the Nonsig criterion. The retention of lower X^2 values as anchors resulted in decreased accuracy of the strategy with a significance value of 0.01. The combination of the AOAA and Nonsig criterion resulted in no contamination within the selected anchor subset. The proportion of replication was recognized as an elevated pure anchor with the help of an alternative criterion for the implementation of AOAA. Higher accuracy was achieved, when the anchor selection was combined with the $MinX^2$ criterion as compared to the NonsigMaxA criterion.

Power and Type I Error of DIF Testing

Table 1 represents the average power and type I error associated with DIF testing. The average power is increased due to the implementation of Wald tests and utilizing AATA-based strategies during the DIF testing; whereas, type I error remains inflated. The type I error remained under the nominal level of 0.05 after the implementation of IRT-LR tests and utilizing anchors selected according to the AOAA strategies. Under the single-anchored conditions, DIF testing possesses a decreased rate of power. The increased risk of contaminated anchor can be out-weighted through the benefits of more designated anchors, but as the number of selected anchors increases, the risk of anchor contamination tends to increase. For instance, if all the items are chosen as anchors, it might result in decreased statistical power. The examination of results based on contamination-free replications depicted the significance of highly discriminated anchors, because of outperformance of NonsigMaxA over $MinG^2$/$MinX^2$.

Table 1: Anchor-selection Accuracy and Subsequent DIF Testing Power and Type I Error

Strategy	Anchor-selection accuracy	DIF power	DIF Type I error	DIF power (with pure anchors)	DIF Type I error (with pure anchors)
MinG2/Minx2					
AATA					
Single	89.31%	0.76	0.16	0.76	0.11
AOAA					
Single	73.22%	0.36	0.00	0.21	0.00
NonsigMaxA					
AATA					
Single	78.77%	0.68	0.23	0.83	0.12
AOAA					
Single	88.21%	0.17	0.00	0.21	0.00

Regression Model and Analysis

The two logistic regression models were used to examine the levels and interactions of anchor items for the optimization of probability to achieve positive outcomes. The results of DIF testing were found satisfactory under such conditions; therefore, the NonsigMaxA conditions have been examined for the respective AATA and AOAA methods. The anchor items have been used in testing the logistic regression models as it helps in acquiring knowledge, which is important for the selection of anchors before conducting DIF analysis.

The study has tested the performance of Wald's AATA-based anchor-selection strategies and made a comparison between different strategies according to the IRT-LR AOAA procedure through simulation processing. As compared to AATA, AOAA possesses increased anchor-selection accuracy under the same conditions (Wang & Woods, 2017). The selection of anchors based on these strategies performed well during the DIF testing and tended to achieve increased statistical power with controlled type I error.

A power of above 0.70 was maintained after the implementation of AOAA-NonsigMaxA strategy; whereas, the type I error was controlled within the range of 0.05. Few past studies have revealed that the IRT-LR depicts extremely low power in the presence of a single item designated as anchor (Wang & Yeh, 2003; Lopez Rivas, Stark, & Chernyshenko, 2009; Meade & Wright, 2012). Moreover, the follow-up DIF tests for the various AATA-based strategies were plagued through the inflated Type I error under different conditions. This is the reason, which does not allow the replacement of AOAA with the AATA procedure in majority of the conditions (Wang & Yeh, 2003; Wang & Woods, 2017). However, Wald test does not need to always perform inefficiently as compared to the IRT-LR test by utilizing the anchors selected by AATA. The Wald test tends to yield high power together with controlled type I error when the anchors are contamination-free (Woods, Cai, & Wang, 2013).

Conclusion

The study revealed new anchor selection strategies and compared them with the existing methods, which are used in DIF analysis. The facilitation and implementation of advance anchor selection strategies have resulted in the straightforward notation of these strategies. The performance of anchor selection strategies, associated with the regression model, has been evaluated with an extensive simulation study. DIF tests results were dependent on the anchor selection strategies employed. Moreover, the results have revealed that appropriate anchor-selection not only depends on the sample size, but it has an association with the proportion of DIF items and the direction of DIF along with the length of the anchor. Test depicts the efficacy of the Wald method for the DIF, which has been corroborated by various other studies (Cao, Tay, & Liu, 2017). The outcomes suggest that anchor method used must be carefully evaluated for avoiding high misclassification and dubious results.

The findings of the study can be generalized, given the conditions adopted for presenting the data in an understandable form. The findings recommend that DIF serve as a meaningful item, where the DIF statistics representation using effect size measures should be used, which help present the dependent and independent variables association (Kirk, 1996; Zumbo, 2007). The findings help guide the practitioners and researchers concerning the selection of the optimal method for DIF detection. This research provides several recommendations, including the DIF testing. It shows that IRT- LR is effective when DIF type prediction, DIF magnitude, and DIF item proportion is difficult. In case these are expected to be above 20 percent, then the Wald test is appropriate.

This study suggests future researchers compare the different modeling approaches for the detection of DIF methods such as multiple indicators multiple causes (MIMIC) modeling following variant simulation techniques. Also, future studies are recommended to analyze the efficiency of the anchor selection strategies concerning different groups and times. As the Wald x 2AATA processes only need model fittings irrespective of the comparison between the parameters sets (Kim, Cohen, & Kim, 1994), it can save computation time as compared to the IRT-LR AOAA model.

References

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An Investigation of the Power of the Likelihood Ratio Goodness-of-Fit Statistic in Detecting Differential Item Functioning. *Journal of Educational Measurement, 36*(4), 277-300. https://doi.org/10.1111/j.1745-3984.1999.tb00558.x.

Bastug, Ö. Y. Ö. (2016). A Comparison of Four Differential Item Functioning Procedures in the Presence of Multidimensionality. *Educational Research and Reviews*, *11*(13), 1251-1261.

Battauz, M. (2019). On Wald tests for differential item functioning detection. *Statistical Methods & Applications*, *28*(1), 103-118. https://doi.org/10.1007/s10260-018-00442-w.

Cao, M., Tay, L., & Liu, Y. (2017). A Monte Carlo study of an iterative Wald test procedure for DIF analysis. *Educational and psychological measurement, 77*(1), 104-118. https://doi.org/10.1177/0013164416637104.

Childs, R. A., Dahlstrom, W. G., & Panter, A. T. (2000). Item response theory in personality assessment: A demonstration using the MMPI-2 Depression Scale. *Assessment*, *7*(1), 37-54. https://doi.org/10.1177/107319110000700103.

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: exploring a new sequential-free baseline procedure. *Applied psychological measurement*, *40*(7), 486-499. https://doi.org/10.1177/0146621616659738.

Hou, L., la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98-125. https://doi.org/10.1111/jedm.12036.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, *72*(3), 469-492.

Kim, S. H., Cohen, A. S., & Kim, H. O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*(3), 217-228. https://doi.org/10.1177/014662169401800303.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, *56*(5), 746-759. https://doi.org/10.1177/0013164496056005002.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer. https://doi.org/10.1007/978-1-4939-0317-7.

Kopf, J., Zeileis, A., & Strobl, C. (2015). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, *39*(2), 83-103. https://doi.org/10.1177/0146621614544195.

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral dissertation, The University of North Carolina at Chapel Hill).

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251-265.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, *28*(4), 989-1020. https://doi.org/10.1177/001316446802800401.

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior research methods*, *42*(3), 847-862. https://doi.org/10.3758/brm.42.3.847.

Marsman, M., Waldorp, L., & Maris, G. (2017). A note on large-scale logistic prediction: Using an approximate graphical model to deal with collinearity and missing data. *Behaviormetrika*, *44*(2), 513-534. https://doi.org/10.1007/s41237-017-0024-x.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*(1), 55-64. https://doi.org/10.1207/s15327906mbr4101_4.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016. https://doi.org/10.1037/a0027934.

Thissen, D. (2001). IRTLRDIF v. 2.0 b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. *Chapel Hill, NC: LL Thurstone Psychometric Laboratory*.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.

Wang, M., & Woods, C. M. (2017). Anchor Selection Using the Wald Test Anchor-All-Test-All Procedure. *Applied Psychological Measurement*, *41*(1), 17-29. https://doi.org/10.1177/0146621616668014.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479-498. https://doi.org/10.1177/0146621603259902.

Wang, W. C., Shih, C. L., & Yang, C. C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, *69*(5), 713-731. https://doi.org/10.1177/0013164409332228.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42-57. https://doi.org/10.1177/0146621607314044.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532-547. https://doi.org/10.1177/0013164412464875.

Zumbo, B. D. (2007). Three generations of DIF analyses: 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.