Anchor Selection Using the Wald Test Anchor-All-Test-All Procedure

Applied Psychological Measurement

2017, Vol. 41(1) 17–29

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621616668014

apm.sagepub.com

Mian Wang¹ and Carol M. Woods²

Abstract

Methods for testing differential item functioning (DIF) require that the reference and focal groups are linked on a common scale using group-invariant anchor items. Several anchor-selection strategies have been introduced in an item response theory framework. However, popular strategies often utilize likelihood ratio testing with all-others-as-anchors that requires multiple model fittings. The current study explored alternative anchor-selection strategies based on a modified version of the Wald χ^2 test that is implemented in flexMIRT and IRTPRO, and made comparisons with methods based on the popular likelihood ratio test. Accuracies of anchor identification of four different strategies (two testing methods combined with two selection criteria), along with the power and Type I error associated with respective follow-up DIF tests, will be presented. Implications for applied researchers and suggestions for future research will be discussed.

Keywords

differential item functioning, anchor selection, Wald's test, anchor-all-test-all, likelihood ratio test, all-others-as-anchors

Differential item functioning (DIF) occurs when the probability of endorsing a given response category differs across groups after individuals are matched on the latent trait (Embretson & Reise, 2000). Some common methods used for DIF detection include the item response theory likelihood ratio test (IRT-LR; Thissen, Steinberg, & Wainer, 1993), the Mantel–Haenszel test (Holland & Thayer, 1988; Mantel & Haenszel, 1959), the logistic regression procedure (Swaminathan & Rogers, 1990), and Lord's Wald χ^2 test and its improved version (Langer, 2008; Lord, 1977, 1980; Wald, 1943). However, all of these methods require a refined group-invariant anchor item/subset to match subjects on the latent factor before carrying out a DIF analysis. Contamination of the anchor item/set (i.e., a group-variant item is mistakenly designated as an anchor) often leads to inflated Type I/II errors and inaccurate parameter estimates (Finch, 2005; Lopez Rivas, Stark, & Chernyshenko, 2009; Stark, Chernyshenko, & Drasgow, 2006; Woods, 2009; Woods, Cai, & Wang, 2013). Therefore,

¹Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA

²The University of Kansas, Lawrence, KS, USA

Corresponding Author:

Mian Wang, 101 E. Weaver St., Suite 220, Carrboro, NC 27510, USA.

Email: mianwang@unc.edu

correct identification of an anchor item/set is a crucial first step before a researcher carries out subsequent DIF analysis.

Literature Review

Anchor-Selection Methods

Several empirical anchor-selection strategies have been introduced and investigated in the past. Probably, the most popular anchor-selection method is the IRT-LR all-others-as-anchors (AOAA) procedure which is a variant of the IRT-LR test (Thissen et al., 1993). The IRT-LR AOAA procedure first fits a baseline model to the data with corresponding reference and focal group item parameters constrained to be equal. Multiple augmented models are then fitted to the data by treating all other items (i.e., all items except the item being analyzed) as anchors, while parameters of the analyzed item are free to vary. Unlike its variant, the original IRT-LR test only constrains a set of known anchors equal between groups in the baseline, and then nested model comparisons are conducted to evaluate the equality constraints placed on each tested item one at a time. Regardless, for both approaches, the actual number of augmented models fitted to the data is at least equal to the number of items being compared between groups. By calculating -2 times the logarithm of the likelihood ratio between each augmented model and the baseline model, the overall test statistic G^2 for each tested item (G^2 follows a χ^2 distribution when the less restrictive augmented model is correctly specified) can be obtained. The G^2 -nonsignificance criterion (or universally for any test statistic, the Nonsig criterion) is then applied where items with nonsignificant G^2 statistics are retained as anchors (Thissen et al., 1993). Note that the actual number of items being selected as anchors depends on the significance cutoff.

Besides the Nonsig criterion, IRT-LR AOAA can be paired with an alternative criterion, such as the $\text{Min}G^2$ criterion (Woods, 2009), or the Nonsig and largest “ a ” combined criterion (i.e., NonsigMaxA; “ a ” refers to an item’s discrimination; Stark et al., 2006).

The $\text{Min}G^2/\text{Min}\chi^2$ Criterion. The $\text{Min}G^2$ criterion (Woods, 2009) selects anchors by ranking items in ascending order (i.e., smallest on top) based on their G^2 values. A small number of items with the smallest G^2 values are then chosen as anchors. The actual number of items being selected as anchors may depend on factors such as test length and sample size per group, although researchers generally suggested not selecting more than 25% out of the total number of items as anchors (Meade & Wright, 2012; Woods, 2009). The $\text{Min}G^2$ approach originated from the assumption that larger G^2 values reflect greater DIF effects. The major distinction of this criterion is that the $\text{Min}G^2$ criterion does not take the (non)significance of the G^2 statistic into account prior to ranking. In the context of a Wald test, the test statistic is named as a χ^2 statistic, and therefore the Wald version of $\text{Min}G^2$ shall be referred to as the $\text{Min}\chi^2$ criterion.

The NonsigMaxA Criterion. For the NonsigMaxA criterion (Lopez Rivas et al., 2009; Stark et al., 2006), only the items with nonsignificant G^2 values are ranked based on their estimated reference group discrimination parameters in descending order (i.e., largest on top), and then a certain number of items (usually not exceeding 25% of the total number of items) with the largest discrimination parameter values are selected as anchors. This approach stems from the notion that the latent trait is better defined by anchor items with higher discrimination (analogous to group-invariant items with higher factor loadings). More importantly, items with higher discriminations were found to be less prone to Type I and Type II errors under the IRT-LR test, especially when the latent distributions of the groups are not identical (see Ankenmann, Witt, &

Dunbar, 1999). In other words, highly discriminating items are more likely to be identified as nonsignificant when they actually are group invariant (i.e., high specificity as the complement of low Type I error), and they are also less likely to be retained as nonsignificant when they exhibit DIF effect (i.e., low Type II error). Given the asymptotic equivalence of the test statistics of the IRT-LR and the Wald tests, it was assumed that these findings are generalizable to results under the Wald test, and the current study evaluated the tenability of this assumption by examining anchor-selection accuracy using the NonsigMaxA criterion under both the IRT-LR and Wald tests.

In a recent simulation study, Meade and Wright (2012) compared the performance of a wide range of anchor-selection strategies and their variants, including but not limited to all of the previously mentioned anchor-selection strategies, by evaluating the power and Type I error of the DIF analysis following each anchor-selection strategy. Their study showed that, in contrast to using IRT-LR AOAA with the Nonsig criterion, the application of an alternative selection criterion (e.g., NonsigMaxA) yielded better results. Even though some other strategies compared by Meade and Wright were promising, the IRT-LR AOAA procedure coupled with an alternative criterion was easiest to implement while maintaining high power and controlled Type I error in the subsequent DIF analysis. Thus, variants of their suggested selection criteria, especially when applied under the Wald test, were of great interest for the current study.

Concerns With the IRT-LR AOAA Procedure

All of the aforementioned anchor-selection techniques depend on the implementation of the IRT-LR AOAA procedure and, despite its popularity, this procedure has some drawbacks. First and foremost, multiple model fittings of the same data are required. For the IRT-LR AOAA procedure, a baseline compact model and multiple augmented models have to be fit. Therefore, the total number of nested model comparisons being conducted can be extremely time-consuming with lengthy tests and large samples. Consequently, there is no easy implementation of IRT-LR AOAA that compares three or more groups due to a massive amount of model fittings required when groups are compared in pairs (using the popular IRTLRDIF software; Thissen, 2001). Moreover, the augmented models used for IRT-LR AOAA are likely misspecified due to potential inclusion of DIF items, so G^2 fails to follow a χ^2 distribution (Maydeu-Olivares & Cai, 2006). Hence, any follow-up tests relying on G^2 cannot be trusted. In addition, the IRTLRDIF software (Thissen, 2001) which made IRT-LR AOAA easily accessible is no longer being maintained or updated and soon will be withdrawn. Therefore, the current study is dedicated to evaluating alternative strategies for anchor selection that could circumvent at least some of the limitations with IRT-LR AOAA.

Lord's Wald χ^2 Test and Its Improved Version

Originally introduced by Lord (1977, 1980) for DIF detection, Lord's Wald χ^2 test compares item parameters between groups using Wald's (1943) χ^2 test statistic. Unlike the IRT-LR test, Lord's Wald χ^2 test in general only requires a single model fitting when testing for DIF. Assuming anchor items are prespecified, the Wald test uses anchors for linking the groups on the same latent scale (subjects' latent scores are obtained), and then parameters of studied items are freely estimated for each group conditioning on their latent scores. Also, Lord's Wald χ^2 statistic is not based on nested model comparisons, so there is no theoretical problem with misspecification of augmented models. Furthermore, this test can be readily extended to DIF analysis of multiple groups through the introduction of a contrast coefficient matrix (see Kim, Cohen, & Park, 1995; Langer, 2008; Woods et al., 2013).

Lord's Wald χ^2 statistic for the joint differences between parameters of two groups is calculated as

$$\chi_i^2 = \mathbf{v}_i^T \Sigma_i^{-1} \mathbf{v}_i,$$

in which $\mathbf{v}_i^T$ holds differences between item parameter estimates, for example, $[\hat{a}_{Fi} - \hat{a}_{Ri}, \hat{b}_{Fi} - \hat{b}_{Ri}]$ for a two-parameter logistic (2PL) model. Σ_i is the sum of the asymptotic standard error covariance matrices associated with the item parameter estimates for both groups, and df equals the number of parameters being compared for each item (i.e., $df = 2$ for 2PL items). An item is flagged as having a DIF effect when its corresponding χ^2 statistic is significant.

The Wald test has recently been improved (Cai, 2015; Cai, Thissen, & du Toit, 2011; Langer, 2008). Specifically, the improved version of Lord's Wald test uses the expectation–maximization marginal maximum likelihood procedure for parameter estimation, the supplemented expectation–maximization (SEM) algorithm for error covariance matrix calculation (Cai, 2008), and the concurrent calibration approach (Kolen & Brennan, 2014) for linking the groups. With these improvements, the current version of the Lord's Wald χ^2 test has been on par with the IRT-LR test in terms of power and Type I error for DIF testing (Woods et al., 2013).

As a variant of the improved version of Lord's Wald χ^2 test, the anchor-all-test-all (AATA) procedure (proposed by Langer, 2008, referred to as “Wald-2” by Woods et al., 2013) requires two model fittings per data set because no anchor items are prespecified. In the first stage, the reference group mean and SD are fixed to 0 and 1, respectively, and then the anchor-all model (i.e., all corresponding item parameters are fixed to be equal between groups; analogous to the baseline model for IRT-LR AOAA) is fitted to estimate the focal group mean and SD . The test-all model is fitted in the second stage where all item parameters are freely estimated with the focal group mean and SD fixed to the values obtained from the first stage. Wald χ^2 test statistics are then calculated based on item parameter differences and the corresponding error covariance matrices. Items with significant χ^2 values are said to exhibit a DIF effect between groups. Woods et al. (2013) found that the Wald χ^2 AATA procedure showed inflated Type I error for DIF detection when implemented as described above. Their study also showed that up to 12% of DIF items were not detected by implementing the Wald χ^2 AATA procedure paired with the Nonsig criterion. However, given the success that the IRT-LR AOAA procedure has achieved with the help of an alternative anchor-selection criterion (Meade & Wright, 2012), it was believed that results would improve when Wald χ^2 AATA is used in conjunction with an alternative criterion.

Current Study

The current simulation study compared four strategies with the cross combinations of two testing methods and two alternative selection criteria: (a) Wald χ^2 AATA paired with the Min χ^2 criterion, (b) IRT-LR AOAA paired with the Min G^2 criterion, (c) Wald χ^2 AATA paired with the NonsigMaxA criterion, and (d) IRT-LR AOAA paired with the NonsigMaxA criterion. These four strategies were first compared in terms of their anchor-selection accuracy (i.e., the proportion of replications that yielded a contamination-free anchor item/subset). The statistical power and Type I error associated with subsequent DIF analysis, using anchors preselected by the corresponding strategy, were also evaluated.

The overarching goal of the current study was to find out whether the IRT-LR AOAA procedure can be replaced by the Wald AATA procedure in previously established anchor-selection

strategies. Conditioning upon satisfactory results, applied researchers could potentially save a considerable amount of time by applying the Wald AATA procedure, because it only requires two model fittings under all conditions. This study also investigated the relationships between the manipulated variables (i.e., test length and sample size) and the DIF test outcomes, revealing the optimal conditions under which the selected anchors would be most likely to yield high power and well-controlled Type I error in the subsequent DIF tests.

Method

Simulation Design

Despite its merits, the traditional factorial design (i.e., factors with fixed levels) for simulation studies is potentially flawed by artificially categorizing continuous variables (e.g., sample size) for the purposes of saving simulation time (data generated only for limited number of levels) and possibly easier summarization of the results (e.g., the conventional factorial analysis of variance approach for analyzing categorized factors). However, with the improved computing speed for simulation and more advanced modeling techniques for summarizing results, researchers should no longer be confined to the standard practice of artificially categorizing continuous factors and risking losing information from the continuous scale.

Therefore, instead of the traditional factorial design approach, the current study adopted a randomized simulation mechanism (i.e., factor levels are randomly drawn from prespecified distributions or ranges of values). Such a design also facilitated the modeling of the results when seeking the optimal conditions for obtaining best outcomes. To diversify the simulated values and improve generalizability of the results, a total of 10,000 replications were conducted for the current study.

Fixed factors. For each replication, there were two groups with equal sample sizes, and all data were generated from a 2PL model. Latent trait levels of reference and focal groups were distributed as $\theta_R \sim (0, 1)$ and $\theta_F \sim N(-0.5, 1)$. The focal group latent mean was lower than the reference group mean to test the anchor-selection strategies' robustness to true impact versus DIF effects, and studies with impact-only conditions were not uncommon in the DIF literature (e.g., Lopez Rivas et al., 2009; Meade & Wright, 2012; Wang & Yeh, 2003; Woods, 2009).

Varying factors. The total sample size was a randomly drawn even integer from a discrete uniform distribution between 400 and 4,000, and therefore each group had a sample size ($n_{Subjects}$) anywhere between 200 and 2,000. The test length (n_{Items}) for each replication was a value randomly drawn from a discrete uniform distribution ranged between 5 and 50. For each replication, there was at least one item with differential functioning properties, and the maximum number of DIF items did not exceed 80% of n_{Items} , so that each replication was guaranteed to have 20% or more group-invariant items from which to select. These boundaries were set so that the resulting data represent a wide range of possible conditions.

With regard to item parameters, reference group difficulty parameters were drawn from $b_R \sim N(0.1, 1.3^2)$ with truncations at -2.5 and 2.4 , and discrimination parameters were drawn from $a_R \sim N(0.9, 0.5^2)$ with truncations at 0.3 and 2.1 . Decisions about item parameter distributions and truncations were informed by 598 dichotomous items from empirical studies within educational and psychological contexts (e.g., Childs, Dahlstrom, Kemp, & Panter, 2000; Lord, 1968; see boldface listings in the reference section for a complete list of studies reviewed). For the magnitudes of DIF effect, item parameter differences were drawn from uniform distributions $|b_{DIF}| \in [0.3, 0.7]$ and $a_{DIF} \in [-0.7, 0.7]$. Theoretically, an item exhibits uniform DIF with only b_{DIF} , and nonuniform DIF when nonnegligible a_{DIF} is also present. Item parameters

of the focal group (i.e., b_F and a_F) were obtained by adding parameter differences (i.e., the magnitudes of DIF effect) to the corresponding reference group item parameters. In other words, with the combination of DIF magnitudes specified above, each DIF item had at least a difference in the location parameters by 0.3 units (in the metric of the standardized latent variable), and a subset of items also differed between groups in terms of their discrimination parameters. Furthermore, any discrimination parameter DIF magnitudes that resulted in zero or negative a_F values were redrawn until all focal group discrimination parameters were positive.

Detailed Steps of the Four Anchor-Selection Strategies

AATA-Min χ^2 . Data were first analyzed using Wald's χ^2 AATA by fitting a baseline model, with all item parameters constrained equal between groups, to link the groups and then estimate the focal group distribution. A full model was also fitted to freely estimate all item parameters, with the focal group mean and *SD* fixed to the values obtained from the baseline model. The differences between the two sets of item parameters of the two groups were tested using χ^2 statistics. Items were then ranked according to their χ^2 values in ascending order. Either a single item or a proportion of items (10% or 20% out of $nItems$) with the smallest χ^2 value(s) was then selected as the anchor item/subset. When the intended number of items being selected ended up with a noninteger value, it was always rounded up to the nearest integer (e.g., $6 \times 10\% = 0.6$ was rounded to 1 when tried to select 10% out of six items as anchors).

AOAA-Min G^2 . The AOAA-Min G^2 strategy was carried out in a similar way as the AATA-Min χ^2 strategy, except that a total of $(1 + nItems)$ number of models were fitted to data in IRTLRDIF (Thissen, 2001), and nested model comparisons were conducted to calculate G^2 test statistics for each item. Items were also ranked according to their test statistics' values in ascending order, and then a fixed number of items (equal to the number of anchors selected for the AATA-Min χ^2 strategy within the same replication) with the smallest test statistics were retained as anchors.

AATA-NonsigMaxA. After acquiring the χ^2 statistics from the Wald AATA procedure (same run as the AATA-Min χ^2 procedure within a replication), only items with nonsignificant omnibus χ^2 values (always screened at $\alpha = .05$) were retained, and then the nonsignificant items were ranked based on their estimated reference group discrimination parameters in descending order. Either a single item or a proportion of items (10% or 20% out of $nItems$) with the largest estimated reference group discrimination parameters was then selected as an anchor item/subset. When the intended number of items being selected ended up with a noninteger value, it was always rounded up to the nearest integer (e.g., $6 \times 10\% = 0.6$ was rounded to 1 when tried to select 10% of six items as anchors). In cases where this strategy failed with all items showing significant χ^2 (thus, no item could be retained for ranking), the Min χ^2 criterion should be applied instead.

AOAA-NonsigMaxA. Based on the G^2 test statistics calculated by the IRT-LR AOAA procedure (same run as the AOAA-Min G^2 procedure within a replication), only items with nonsignificant G^2 values were kept for the follow-up ranking. A prespecified number of nonsignificant items (also the same as other procedures) with the largest estimated reference group discrimination parameters were identified as an anchor item/subset.

Follow-Up DIF Analysis

For each replication, a total of four follow-up DIF tests were carried out, respectively, using the anchor set preselected by its corresponding anchor-selection strategy. The one-stage improved

Wald test was implemented for DIF detection when an anchor item/subset was provided by an AATA-based strategy, whereas the traditional IRT-LR test was used in conjunction with the AOAA-based strategies.

Outcome Evaluation

The accuracy associated with each anchor-selection strategy was evaluated by calculating the percentage of replications that had identified a pure anchor item/subset out of the 10,000 replications, under each of the simulation conditions.

The average power (i.e., true positives) and Type I error (i.e., false positives) associated with the follow-up DIF analysis were also reported. For each replication, power was the percentage of true DIF items being correctly flagged as having a DIF effect, and Type I error was the percentage of DIF-free items being mistakenly flagged as having a DIF effect. In situations where an anchor-selection strategy failed to provide the required number of anchors for its follow-up DIF testing (e.g., in cases where all items were flagged as significant by AATA/AOAA, and thus the NonsigMaxA criterion was not applicable), the average power and Type I error were calculated with those cases excluded.

Finally, for the best performing condition of each underlying test (AATA or AOAA), a dichotomous variable was created to indicate whether a replication had achieved the ideal power ($\geq .80$) and Type I error ($\leq .05$) during DIF testing, and logistic regression was used to examine the effect of $nItems$, $nSubjects$, and their interaction on this dichotomous outcome.

Software

Wald's χ^2 AATA was implemented in flexMIRT version 3 (Cai, 2015), and IRT-LR AOAA was carried out using IRTLRFID version 2 (Thissen, 2001). R version 3 (R Core Team, 2015) base package was used for data generation, flexMIRT and IRTLRFID scripts compilation, output collection, and results summarization.

Results

Focal Group Estimation

Because the second stage of Wald's χ^2 AATA estimates item parameters based on fixed focal group mean and SD (estimated in the first stage), the bias in estimation of the focal group distribution will be reported here. Across all replications, the average estimated focal group mean was -0.34 (the true focal group mean was -0.5) with a range between -1.05 and 0.46 , whereas the average estimated focal group SD was 0.83 (the true focal group SD was 1) with a range from 0.61 to 1.12 . The root mean squared errors for the focal group mean and SD estimates were 0.24 and 0.18 , respectively. With Wald's χ^2 AATA, such biased estimation of the focal group distribution was not surprising given that the first stage of the procedure used all items (including DIF items) as anchors for estimating the focal group distribution.

Overall Accuracy of Anchor Selection

To help rule out potential simulation errors, accuracy associated with the naïve way of picking anchors using the Nonsig criterion on top of the AATA or AOAA procedure was calculated first. As expected, AATA combined with only the Nonsig criterion had extremely low accuracy

(31.26% with $\alpha = .05$). The accuracy of this strategy (with $\alpha = .01$) dropped to 21.16% as more items with lower χ^2 values were retained as anchors. In contrast, 40.57% and 28.31% of the selected anchor item/subset had no contamination, using AOAA combined with the Nonsig criterion at $\alpha = .05$ and $\alpha = .01$, respectively.

Despite the Nonsig criterion's disappointing performance, with the help of an alternative criterion, the proportion of replications that identified a pure anchor item/subset markedly increased, especially when AOAA was implemented. The anchor-selection accuracy associated with the $\text{Min}\chi^2$ criterion was always higher than that with the NonsigMaxA criterion, no matter which underlying test (AATA or AOAA) was implemented and/or how many items were intended to be selected as anchors. See the first column of Table 1 for more details.

Overall Power and Type I Error of the Subsequent DIF Tests

The average statistical power and the average Type I error are, respectively, shown in the second and third columns of Table 1. During the actual DIF tests, when the one-stage improved Wald tests were implemented using anchors preselected by AATA-based strategies, the average statistical power had always stayed above .75, whereas the average Type I error was inflated with a range between .09 and .23. As to the IRT-LR tests implemented using anchors preselected by AOAA-based strategies, the average Type I error was unanimously under the nominal level of .05, although the DIF tests had extremely low power under the single-anchor conditions.

Overall, the conditions with 20% as anchors had better power than the conditions using 10% as anchors, despite that the 20%-as-anchors conditions had more anchor contamination issues. The benefit of more designated anchors outweighed the increased risk of a contaminated anchor set, at least among the conditions evaluated in the current study. However, this benefit is expected to vanish as the percentage of items being selected as anchors increases (e.g., statistical power will drop to nil if all items are chosen as anchors).

In addition, even though the $\text{Min}G^2/\text{Min}\chi^2$ conditions were expected to outperform the NonsigMaxA conditions given the anchor-selection accuracies shown in Table 1, results from the subsequent DIF tests suggested otherwise. It appeared that having a set of highly discriminating anchors (with possible contamination) was more crucial than ensuring a pure anchor set. The importance of having highly discriminating anchors was further confirmed by examining the results based on only contamination-free replications, as NonsigMaxA outperformed $\text{Min}G^2/\text{Min}\chi^2$ under all conditions except one (see the fourth and fifth columns of Table 1).

Logistic Regression Models

As mentioned earlier, for the two conditions where 20% of items were chosen as anchors using either AATA-NonsigMaxA or AOAA-NonsigMaxA, dichotomous outcome variables were created to indicate whether a replication had achieved the ideal power (at or above .80) and Type I error (at or below .05). Two logistic regression models were then fitted to examine the levels of $n\text{Items}$, $n\text{Subjects}$, and their interaction that would optimize the probability of achieving the ideals, when anchors (20%) were preselected by either of the two strategies.

Only the intersection of 20%-as-anchors and NonsigMaxA conditions was examined for each respective base method (AATA/AOAA), because the DIF test results found under such conditions were the most promising. Only $n\text{Items}$ and $n\text{Subjects}$ were used as the predictors in the logistic regression models, because they are likely the only pieces of information that a researcher would know prior to selecting anchors for a DIF study.

Table 1. Accuracy of Picking an Anchor Item/Subset Without Contamination, and Subsequent DIF Tests' 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/Minχ^2					
AATA					
Single	79.47%	.76	.23	.79	.17
10%	61.13%	.77	.18	.81	.10
20%	48.18%	.77	.14	.84	.06
AOAA					
Single	91.56%	.21	.00	.22	.00
10%	81.99%	.60	.03	.62	.03
20%	70.09%	.73	.05	.77	.04
NonsigMaxA					
AATA					
Single	74.67%	.78	.22	.81	.12
10%	54.83%	.77	.14	.83	.06
20%	41.64%	.76	.09	.85	.04
AOAA					
Single	88.21%	.19	.00	.20	.00
10%	76.80%	.70	.04	.73	.03
20%	64.72%	.78	.05	.83	.04

Note. For the NonsigMaxA strategy, the significance cutoff was $\chi^2_{0.05}(2) = 5.99$. Items with χ^2 values smaller than the cutoff were retained as nonsignificant items for ranking. For the subsequent DIF tests, the one-stage improved Wald test was implemented using anchors picked by the AATA-based strategies, and the IRT-LR test was implemented using anchors picked by the AOAA-based strategies. DIF = differential item functioning; AATA = anchor-all-test-all; AOAA = all-others-as-anchors; IRT-LR = item response theory likelihood ratio test.

Note that a change of one subject seemed trivial for real-world applications; hence, every 100 subjects was considered as a unit when interpreting the two logistic regression models (i.e., coefficients associated with *nSubjects* were multiplied by 100 before exponentiation).

AATA-NonsigMaxA (20%). A significant interaction effect between *nItems* and *nSubjects* was found, and all coefficients were converted into their exponential forms: $e^{\beta_{nItems}} = 1.011$, $p = .014$ (95% CI = [1.002, 1.020]), $e^{100 \times \beta_{nSubjects}} = 1.179$, $p < .001$ (95% CI = [1.153, 1.206]), and $e^{100 \times \beta_{nItems \times nSubjects}} = 0.998$, $p < .001$ (95% CI = [0.998, 0.999]).

Therefore, the authors of the present study could not interpret the main effects directly, and instead they focused on the interaction between *nSubjects* and *nItems*.

On one hand, with every additional item in the test, the model predicted a 0.79% increase in the odds of achieving the ideal power and Type I error, when sample size was held at 200 subjects per group. The change in odds reversed to a 2.19% decrease with every unit increase in test length, when sample size was held at 2,000 subjects per group. Thus, the effect of test length on the outcome varied at different levels of sample size. Specifically, *nItems* changed the direction of its effect on the outcome at around 670 subjects per group.

On the other hand, when test length was held constant at five items, the odds of achieving the ideal power and Type I error increased by 16.92% for every 100 additional subjects per group. With a 50-item test, the odds only increased by 8.47% for every 100 additional subjects. In conclusion, having a longer test diminished the positive impact of *nSubjects* on the outcome.

In conclusion, when the AATA-NonsigMaxA strategy was used for selecting 20% of items as anchors, the predicted maximum probability (i.e., .70) of achieving the ideal power and Type I error was reached with test length minimized but sample size maximized (i.e., five items and

Figure 1. Probability of achieving the ideal power ($\geq .80$) and Type I error ($\leq .05$).

Note. AATA = anchor-all-test-all; AOAA = all-others-as-anchors.

2,000 subjects in the current simulation). See Figure 1 (left panel) for a three-dimensional (3D) logistic surface associated with this model.

AOAA-NonsigMaxA (20%). A significant interaction effect was also found between $nItems$ and $nSubjects$, and the exponentiated coefficients were $e^{\beta_{nItems}} = 1.027, p < .001$ (95% CI = [1.019, 1.036]), $e^{100 \times \beta_{nSubjects}} = 1.158, p < .001$ (95% CI = [1.134, 1.182]), and $e^{100 \times \beta_{nItems \times nSubjects}} = 0.999, p = .043$ (95% CI = [0.999, 1.000]).

On one hand, the model predicted a 2.57% increase in the odds of achieving optimal DIF test results with every additional item in the test when sample size was held at 200 subjects per group. The change in odds reduced to a 1.31% increase per every additional item when sample size was held at 2,000 subjects per group. Thus, the effect of test length on the outcome diminished slightly as sample size increased.

On the other hand, when test length was held constant at five items, the odds of achieving optimal DIF test results increased by 15.38% for every 100 additional subjects per group. With a 50-item test, the same odds increased by 11.87% for every 100 additional subjects. In conclusion, having a longer test also slightly reduced the positive impact of $nSubjects$ on the outcome.

In conclusion, when the AOAA-NonsigMaxA strategy was used for selecting 20% of items as anchors, the predicted maximum probability (i.e., .73) of achieving the ideal power and Type I error was reached with both test length and sample size maximized (i.e., 2,000 subjects and 50 items in the current simulation). See Figure 1 (right panel) for a 3D logistic surface associated with this model.

Discussion

The current study was the first simulation to test the performance of Wald's AATA-based anchor-selection strategies and to make comparisons with strategies based on the IRT-LR AOAA method. In general, AOAA always had higher anchor-selection accuracy than AATA under the same conditions. Nevertheless, probably the most important question is whether the anchors selected by these strategies performed well in the subsequent DIF tests in terms of achieving high statistical power and well-controlled Type I error.

As discussed in the foregoing sections, AOAA-NonsigMaxA was the only strategy that was able to maintain power above .70 while having Type I error controlled under .05. However, IRT-LR had extremely low power when only a single item was designated as the anchor, and this finding was consistent with past research (Lopez Rivas et al., 2009; Meade & Wright, 2012; Wang & Yeh, 2003; Woods, 2009).

As for the AATA-based strategies, their follow-up DIF tests were plagued by inflated Type I error under all conditions. Therefore, the AATA procedure would not be an ideal replacement for AOAA in most circumstances. However, this is not to say that the one-stage Wald test (using anchors selected by AATA) would always perform worse than the IRT-LR test. In fact, when the anchors had no contamination, Wald yielded high power along with controlled Type I error, especially when 20% of contamination-free anchors were available (see the fourth and fifth columns of Table 1; see also Woods et al., 2013, for similar results).

Limitations and Future Directions

The current study only focused on dichotomous data generated from the 2PL model. Future endeavors should shed light on the performance of the anchor-selection strategies using other types of data (e.g., three-parameter logistic model). For simplicity, the sample sizes for the reference and focal groups were also fixed to be equal. Future research needs to investigate conditions of unequal sample sizes because there are situations where the sample sizes are unequal. In addition, the type of DIF effect (uniform vs. nonuniform) was not monitored in the current study, although such a factor could potentially impact the outcomes investigated.

A possible extension to the current line of research is to study the performance of these anchor-selection strategies in the context of multiple groups or multiple time points (i.e., longitudinal data). Because the Wald χ^2 AATA procedure only requires two model fittings regardless of the number of parameter sets being compared (although a contrast coefficient matrix is required; see Kim et al., 1995), it should be able to conserve a considerable amount of computation time compared with the IRT-LR AOAA procedure.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Boldface entries were used for the review of empirical item parameters.
- Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM criteria for major depression: Evaluating symptom patterns using latent-trait item response models. *Psychological Medicine*, 35, 475-487. doi:10.1017/S0033291704003563
- Allison, K. C., Engel, S. G., Crosby, R. D., de Zwaan, M., O'Reardon, J. P., Wonderlich, S. A., . . . Stunkard, A. J. (2008). Evaluation of diagnostic criteria for night eating syndrome using item response theory analysis. *Eating Behaviors*, 9, 398-407.
- 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, 277-300.

- Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology*, 61, 309-329. doi:10.1348/000711007X249603
- Cai, L. (2015). flexMIRT® 3.0: Flexible multilevel and multidimensional item response theory analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.
- Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International.
- Childs, R. A., Dahlstrom, W. G., Kemp, S. M., & Panter, A. T. (2000). Item response theory in personality assessment: A demonstration using the MMPI-2 depression scale. *Assessment*, 7, 37-54.**
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, 29, 278-295. doi: 10.1177/0146621605275728
- Gomez, R., Vance, A., & Gomez, A. (2011). Item response theory analyses of parent and teacher ratings of the ADHD symptoms for recoded dichotomous scores. *Journal of Attention Disorders*, 15, 269-285. doi:10.1177/1087054709356404**
- Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26, 3-24.**
- Holland, W. P., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.
- Kim, S., Cohen, A. S., & Park, T. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, 32, 261-276.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer Science + Business Media. doi:10.1007/978-1-4939-0317-7
- Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Unpublished doctoral dissertation). University of North Carolina, 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, 251-265. doi:10.1177/0146621608321760
- 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.**
- Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam, The Netherlands: Swets & Zeitlinger.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using $G^2(\text{dif})$ to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, 41, 55-64. doi: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, 1016-1031. doi:10.1037/a0027934
- R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from <http://www.R-project.org/>
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology*, 91, 1292-1306. doi: 10.1037/0021-9010.91.6.1292

- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370. doi:10.1111/j.1745-3984.1990.tb00754.x
- Thissen, D. (2001). IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. Documentation for computer program [Computer software and manual]. Chapel Hill: L.L. Thurstone Psychometric Laboratory, University of North Carolina.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-111). Hillsdale, NJ: Lawrence Erlbaum.
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54, 426-482.
- Wang, W., & Yeh, Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, 27, 479-498. doi: 10.1177/0146621603259902
- Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, 33, 42-57. doi: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, 532-547. doi:10.1177/0013164412464875