GMAC®

Graduate Management Admission Council®
Creating Access to Graduate Business Education®

Differential Impact as an Item Bias Indicator in CAT and Other IRT-based Tests

Fanmin Guo, Lawrence M. Rudner, & Eileen Talento-Miller

GMAC® Research Reports • RR-06-09 • July 17, 2006

Abstract

Differential item functioning methods are widely used in linear tests for detecting potentially biased items, based on pair-wise comparisons between focal and reference groups. In this paper, the authors redefine item bias for computer adaptive tests (CAT) using the perspective of adverse impact.

In item response theory (IRT)-based CAT tests, pre-calibrated and scaled operational item parameters are used for selecting items and then scoring the tests. Item bias has been defined as the difference between the Item Characteristic Curves (ICC) across subpopulation groups. The focus in this paper is whether the use of the operational item parameters is fair to a subpopulation group. Impact is defined as the difference between the ICC of a group and the ICC defined by the operational item parameters. More specifically, an item is biased if examinees in a subpopulation group with a given ability do not have the same conditional probability of correct answers as those with the same ability in the population (all groups combined in calibrating the operational item parameters). If this is not true, the group will be impacted more positively or negatively. Test fairness might be established if none of the items administered has differential impact on the subpopulation groups.

Statistics for flagging differential item impact (DII) are discussed with an example from a U.S.-based CAT test, the Graduate Management Admission Test®. This method also applies to other IRT-based tests.

Introduction

Differential item functioning (DIF) methods are widely used for detecting potentially biased items. The DIF methods compare the performance of a focal group, such as non-native English speakers, with a reference group, such as native English speakers. An item that shows differential performance between the two groups by examinees of the same ability is flagged by the DIF methods for further review. If the differences in group performance are due to skills or knowledge that are not assessed by the test, the flagged items are considered biased and thus either removed from the tests before equating and scoring or balanced out with other items on the same tests. The DIF analyses are repeated for each focal group compared with the reference group for all individual items. All current methods share a common design of comparing two groups at a time after matching examinees on their ability levels, which can be based on scaled scores, number-right raw scores, or theta (θ) estimates.

Under the framework of the item response theory (IRT), item parameters are first calibrated, scaled to a reference scale, and then used to estimate the examinees' abilities as θ estimates. The θ estimates are transformed to a reporting scale before scores are reported. For IRT-based computer adaptive tests (CAT), where the test is tailored to the examinee, the scaled item parameters are used both in selecting items and in estimating the ability of the examinee. It is obvious that the crucial link between the test items and the examinees' scores is the operational item parameters. Unfortunately, IRT-based DIF procedures have followed the same design of their non-IRT counterparts and have not studied this crucial link and, thus, have not taken full advantage of IRT. In this paper,

the authors will propose a different method for identifying potentially biased items under IRT framework using the perspective of adverse impact.

IRT Differential Item Function Methods

An important characteristic of IRT is the item parameter invariance (Lord, 1980). If the unidimensional assumption of the test is met, an item response function or item characteristic curve (ICC) defined by its item parameters will remain unchanged across subpopulation groups. An ICC estimated from any group will be equal to an ICC from another group, and both will be equal to the ICC estimated from responses of all examinees. Lord provided a theoretical basis for both the IRT DIF methods and our proposed method for evaluation of a differential impact.

In the past 25 years, IRT-based DIF statistic methods have been developed and employed to flag potential differential functioning items. Most of the IRT DIF procedures involve three major components:

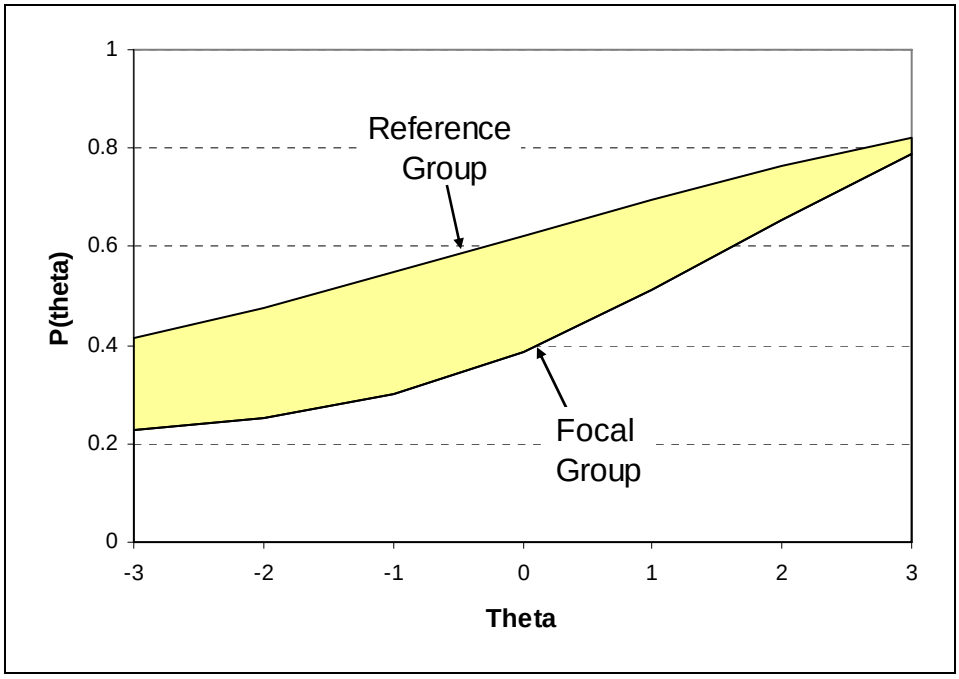
1. Estimate two sets of item parameters—one for the reference group and the other for the focal group. The item parameters are calibrated from the data of a separate group and then put on the same scale.
2. Estimate the difference between the ICCs defined by the two sets of item parameters.
3. Either standardize the differences or test the equity of the two ICCs with statistical methods.

If the standardized difference is small or the statistical test is non-significant, the item is considered unbiased against or in favor of the focal group.

An example of the IRT DIF method is the area index by Rudner (1997) and Rudner, Getson, and Knight (1980). It measures the area between the two ICCs of the reference and the focal groups as an index of the difference between the performances of the two groups matched on ability. The index can be computed either with the signs or without the signs for the differences. The difference can also be presented graphically (See Figure 1). The larger the area, the larger the difference is between the two groups. This is a measure of raw difference. Other methods also involve standardizing this difference or testing the equity of the two ICCs. For a list of the IRT DIF methods, see Camilli and Shepard (1994).

A disconnect exists between the IRT DIF analyses and the item parameters used operationally. The comparison is made between the two group ICCs, each being defined by its group item parameters. However, both sets of item parameters are discarded after the DIF analyses and a new set of item parameters for the item is calibrated and scaled using the examinee data from all groups. It is this new set of item parameters that is used in test assembly, test form equating, and estimation of examinee ability. In this paper, we will refer to the new item parameters estimated from examinees of all groups and used in the test operations as "operational item parameters."

Figure 1. Area Representing Difference between Focal and Reference ICCs



It is important to note that an implicit assumption is made when the IRT DIF methods are employed. That is, the operational item parameters will always show DIF if one of the pair-wise group comparisons shows DIF. This assumption might not always be true. For example, when the population is divided into two mutually exclusive groups, such as the male vs. female or the U.S. vs. non-U.S. groups, the operational ICC for an item is a combination of the two group ICCs. As such, the magnitude of difference between the operational ICC and either group ICC might not be as striking as the difference between the ICCs of the two groups. When there are more than two groups, such as for native language, there may be no difference detected between the ICCs for a particular focal and reference group, though conceivably both might differ from the operational ICC. Another example is a CAT test. In the context of a computer adaptive test, the accuracy of the group ICCs due to small sample size might also present problems for the DIF studies. When new items are pretested for calibrations before they are used in CAT tests, they are usually given to a limited number of random examinees for security reasons. Small sample size will lead to unstable group ICCs, which might lead to false flagging of items for DIF. In each of these situations, the viability of the implicit assumption is questionable. Rather than trying to evaluate the above assumption, the authors of this paper will propose a method for identifying potentially biased items under IRT framework using the perspective of adverse impact.

Differential Item Impact

Since the operational item parameters are used in test assembly, equating, and/or scoring, the evaluation of potential bias of an item against a group should be made by comparing the operational ICC of the item and the performance of a group. An item is not biased against or in favor of a group if the $P(\theta)$ for a group is equal to the $P(\theta)$ for all examinees. In other words, the conditional probability of a correct answer of a particular group is equal to that for all groups combined. If this is not true: $P_g(\theta) \neq P_o(\theta)$, the group in question might be impacted by the use of operational item parameters more negatively or positively than the test population. What is important here is that there should be no differential item impact (DII) among the groups when the operational item parameters are used in test assembly, equating, and/or scoring. An item is flagged as a DII item if $P_g(\theta) \neq P_o(\theta)$. Therefore, the fairness of a test is evident when it includes no DII items or items that differentially impact

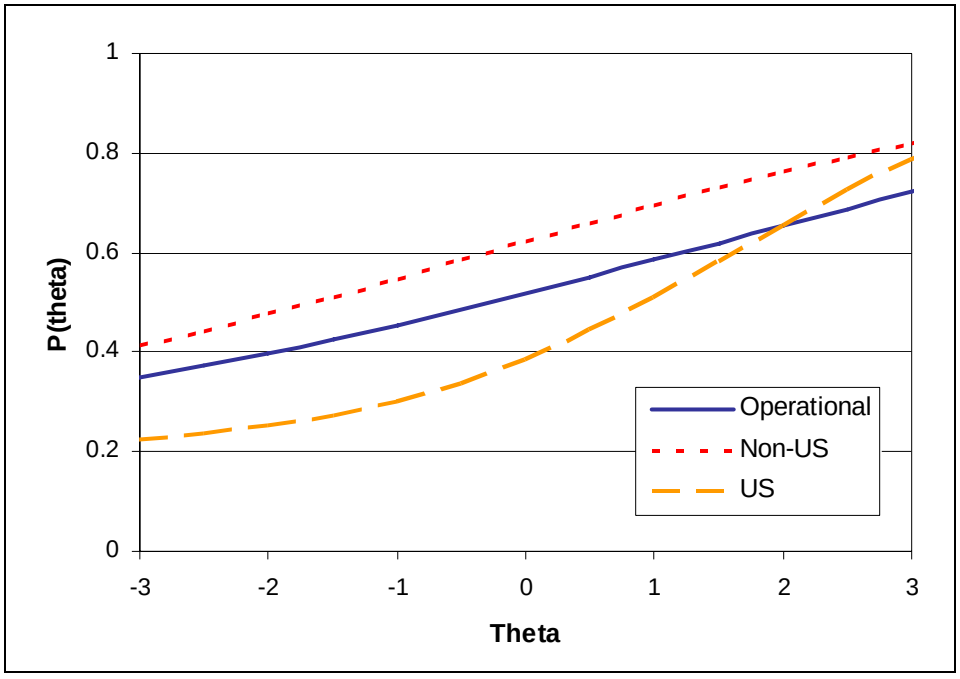
subpopulation groups. Two families of statistical methods lend themselves to identifying potential DII items.

Borrowing IRT DIF Methods for DII

If the sample sizes for estimating group ICCs are large enough to guarantee stable group item parameters, most IRT DIF methods can be used to identify potential DII items. However, the reference group ICCs will be replaced by the operational ICC in each of the analyses. This way, the individual group ICCs are always compared with the operational ICC of an item. An example is given in Figure 2. The examinee responses are real data from a verbal item written for the Graduate Management Admission Test® (GMAT®). The item has never been used in the test due to concerns over its statistical properties. The data were collected from a random sample of GMAT® examinees. Because the θ estimate for each examinee is known from their GMAT CAT® test, the three ICCs in Figure 2 were estimated using a fixed-θ calibration method with the 3-parameter logistic (3PL) IRT model. The resultant ICCs are all on the same θ scale.

In Figure 2, the navy solid line is the operational ICC estimated with data from all examinees (n = 2277); the red dotted line is the ICC for the non-U.S. examinees (n = 1039); and the orange dashed line is the ICC for the U.S. examinees (n = 1238). The operational ICC is always lower than the non-U.S. ICC, indicating that the conditional probability of a correct answer for the non-U.S. examinees is higher than that of the population across the entire θ scale. If the difference is large enough to cause concerns of fairness, it will be flagged as a DII item. This item might also be flagged as a DII item for U.S. citizens because the operational ICC is also higher than the U.S. ICC for most of the θ scale, indicating that the conditional probability of a correct answer for the U.S. examinees is lower than that of the population for most of the θ scale. If the operational item parameters of the item had been used in computer adaptive GMAT® administrations, it might have exerted differential impacts on both groups in item selection and ability estimation.

Figure 2. Operational and Group ICCs of a Test Item



Linn and Harnisch (1981) proposed a DIF procedure for situations where the sample size of the focal group is small. It was later named "Pseudo-IRT Z" by Shepard, Camilli, and Williams (1985). The method starts with estimating an ICC using data from both focal and reference groups. Then the reference group members are divided into quintile intervals based on their abilities. A standardized difference between the ICC and focal group performance are calculated for each quintile interval.

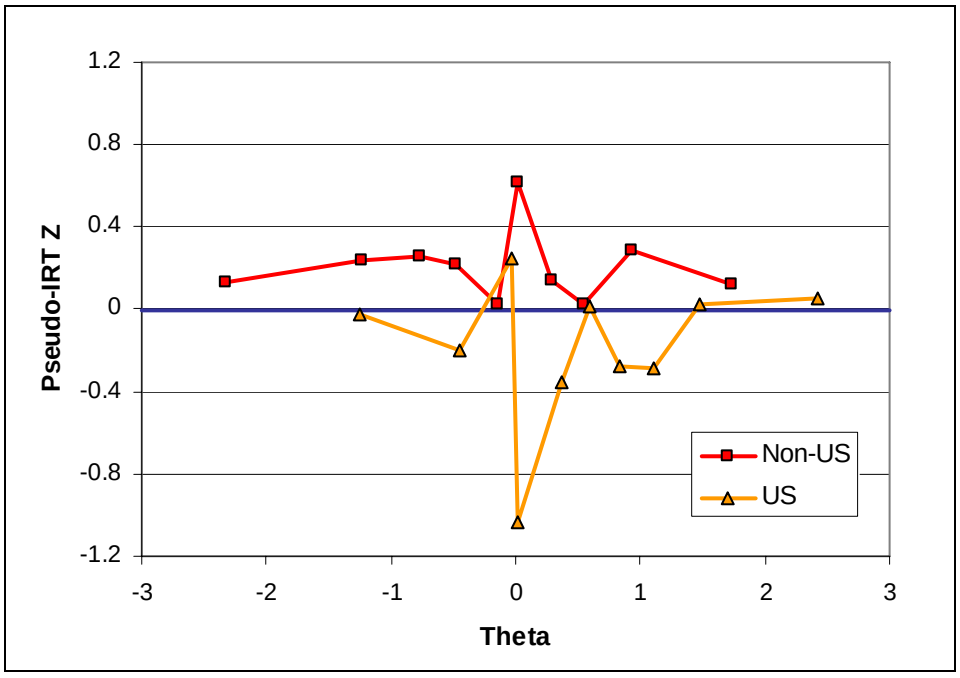
$$Z_{iq} = \frac{1}{n_q} \sum_{j \in q} \frac{U_{ij} - P_{ij}}{\sqrt{P_{ij}(1-P_{ij})}},$$

where i is an item, q is the quintile interval, j is the examinee, U_{ij} is the scored response of an examinee (1 for a correct answer and 0 for a wrong answer), and P_{ij} is the probability of a correct answer for the examinee given his or her $\hat{\theta}$. An overall index is computed by averaging across the ten quintile intervals.

$$Z_i = \frac{\sum_q n_q Z_{ij}}{\sum_q n_q}$$

The pseudo-IRT Z can be used for our purposes if the P_{ij} is replaced with the operational ICC. Figure 3 is an example from the same GMAT® verbal item described above. It is a plot of two pseudo-IRT Z analyses, the standardized residuals from the operational ICC, one for the non-U.S. and the other for the U.S. examinees. For the convenience of presentation, both are plotted on the same graph. For the non-U.S. group, the standardized differences from the operational ICC are all positive, indicating that their conditional probabilities of a correct answer are higher than those predicted by the operational ICC. For the U.S. group, the opposite is mostly true.

Figure 3. Pseudo-IRT Zs by Quintile Group for Non-U.S. and U.S. Examinees



The overall pseudo-IRT Z is .2065 for the non-U.S. group and –.1845 for the U.S. group. Given the magnitude of the standardized differences, this item would be flagged twice as a potential DII item and would be sent to the item writers for content and fairness reviews.

Identifying DII as Model-Data Misfit

In Figure 2, the operational ICC is lower than both the non-U.S. and U.S. ICCs in the θ range of 2 to 3. This might result from the differences between the a-parameter of the operational ICC and the a-parameters of the group ICCs. What this means in terms of item impact is that the conditional probability of a correct answer defined by the operational parameters is lower than those of both groups. A serious concern may be raised if the estimation of the group a-parameters is not accurate because of the small group sample sizes. Generally, a larger sample size is needed in order to obtain stable a-parameters compared with just calibrating b-parameters. Instead of estimating the item parameters for subpopulation groups and comparing them with an operational ICC, evaluating how well the operational ICC fits the observed group data will be a useful tool for identifying potential DII items.

If a test item is fair to members of a group, its operational ICC will correctly predict the conditional probability of a correct answer for a group. That means the operational ICC fits the data from all examinees and it also fits the data from each of the groups in question. DII can now be defined as the misfit of the operational ICC to the data from a group. It should be pointed out that all group misfit does not result from DII items, but DII could be one of the reasons that the operational ICC does not fit the group data. Fit statistics have been in wide use to exclude items for which item parameters cannot be reasonably estimated or for which the estimated ICCs are quite different from their observed data. In our case, it is a special use of the current fit statistics to evaluate the fit between the operational ICC and the data from a group. Yen's Q_1 (Yen, 1981) fit statistic is often used for evaluating fit by studying how well a model could predict the observed data. As an example, we replaced the expected proportion with $P(\theta)$ from the operational ICC and applied it to the same GMAT® item. Yen's Q_1 statistic is defined as

$$Q_1 = \sum_{j=1}^{m} \frac{N_j[P_j - E(P_j)]^2}{E(P_j)[1-E(P_j)]},$$

where j is the ability groups: 1, 2, … , m; P_j is the observed proportion of correct answers in an ability group j; $E(P_j)$ is the expected proportion of a correct answer in j and can be computed as $P(\theta)$ from the operational ICC; and N_j is the number of examinees in j.

Figure 4. Operational ICC and Observed Data

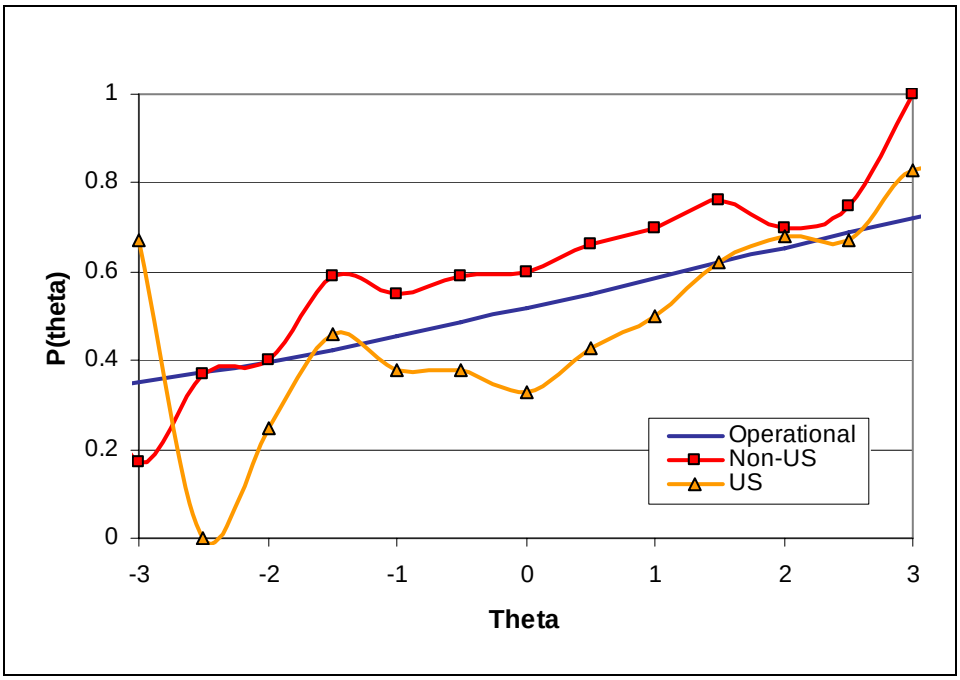


Figure 4 is the operational ICC and the conditional probabilities of a correct answer for the non-U.S. and U.S. examinees separately. Again, the two separate analyses of the two groups are plotted on the same graph. The θ scale between -3 to 3 is divided into 13 equal intervals, and the proportion of correct answers in each interval is calculated and plotted for both groups. It is obvious that most of the observed differences are positive for the non-U.S. examinees and negative for the U.S. examinees. Q_1 for the non-U.S. group is 48.62, and Q_1 for the U.S. group is 66.32. The degrees of freedom for both significance tests are 13 – 3 = 10, and the critical value associated with the df = 10 is 15.99 at the .01 level. Therefore, the operational ICC fits neither of the group data. This item would be flagged by each significance test as a potential DII item.

Conclusions and discussions

This paper makes a distinction between differential item function, DIF, across two different subpopulation groups and differential item impact, DII, the impact of the operational item parameters on a single group. The basic difference is the reference group. In DIF analysis, the reference group is another subpopulation group, typically a majority subpopulation. In DII analysis, the reference is the operational ICC represented by the operational parameters. This distinction is not trivial. DIF examines one group relative to another. DII flags items that appear to affect performance of a group. As such, DII addresses the question of greater concern to the test developer. Is this item appropriate for examinees from each focal group?

In this paper, methods for identifying DII items are discussed and demonstrated using an item written for the GMAT CAT® exam. Pseudo-IRT Z, a traditional DIF method, and Yen's Q_1, a fit statistic, were used with minor modifications on example item. They both flagged the item as having potential DII for the two groups examined. Further review of this item, and any item likewise identified as potentially impacting group performance differently, will help to insure the fairness of the test so that examinees of the same ability have equal probability of success regardless of group membership.

The illustration provided here showed a single item comparing two mutually exclusive groups. As shown in Figure 2, the distance between curves across most of the ability levels was larger between the two groups than between either group compared to the population, thus this item would also be flagged using the more traditional DIF methods. It is important to reiterate that DII and DIF might lead to different conclusions on whether to flag an item. If, in a 3-group case, both the reference and focal group ICCs were under the total group ICC, there may be little evidence for DIF for the focal group. However, the distance from the operational ICC for the focal group may still warrant further examination for DII.

Although great efforts are made to insure tests are fair for all examinees, in reality evaluation of items can only be done for some subpopulation groups. This paper compares U.S. to non-U.S. examinees, but the non-U.S. examinee group is made up of numerous nationalities which, when lumped together, may conceal true differences among different groups that are not being recognized. Both methods demonstrated here can be used with smaller group sizes than would be necessary to calculate stable parameters for ICC curves for each group. Therefore, these methods can be used to evaluate DII among groups with smaller sample sizes to avoid grouping together several focal groups who may not be similar.

Though DII was demonstrated using an example from an IRT-based CAT-based exam, the logic extends to linear IRT tests and tests not using IRT. The key question is whether the use of the operational item parameters, be they 3PL IRT parameters or simply conditional p-values, is appropriate for each focal group. Since DIF methods are well established and routinely used, future research might compare DIF and DII analyses via simulations or actual test administrations. Differences in the nature of the items that are identified would be illuminating. One can conjecture that DII will identify truly problematic items that would not be identified by a DIF analysis.

Contact Information

For questions or comments regarding study findings, methodology or data, please contact the GMAC Research and Development department at research@gmac.com.

Acknowledgements

This paper was presented at the International Test Commission's 5th Conference, July 6-8, 2006, Brussels, Belgium.

References

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. California: Sage.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum.

Linn, R., & Harnisch, D. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*(2), 109-118.

Shepard, L., Camilli, G., & Williams, D. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*(2), 77-105.

Rudner, L. (1977). *An approach to biased item identification using latent trait measurement theory*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Rudner, L., Getson, P., & Knight, D. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5*, 213-233.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.