Journal of Computerized Adaptive Testing

Volume 9 Number 2

November 2022

The (non)Impact of Misfitting Items in Computerized Adaptive Testing

Christine E. DeMars

DOI 10.7333/2211-0902008

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

www.iacat.org

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 9, Number 2, November 2022

DOI 10.7333/2211-0902008

ISSN 2165-6592

The (non)Impact of Misfitting Items in Computerized Adaptive Testing

Christine E. DeMars

James Madison University

To assess the potential impact of misfitting items, simulees received varying percentages of misfitting items. The fit was manipulated to be poor near what would otherwise be the point of maximum information. With 30% misfitting items, the absolute value of the bias of the θ estimates tended to be larger than it was with 0% or 10% misfitting items. However, the magnitude of this effect was small. For most θs and test lengths, the empirical standard error did not vary greatly with the percentage of misfitting items. The standard error estimated from the information function tended to underestimate the empirical standard error when there were 30% misfitting items, but only for higher θ levels. Overall, the misfit had little practical impact.

Keywords: item fit, three-parameter logistic model, computerized adaptive testing

When item response theory (IRT) is used in scoring or developing a test, items that do not fit the model might impact score estimates and item misfit might affect computerized adaptive testing (CAT) more than in fixed-item tests. In CAT, the scale score is typically a direct linear transformation of the IRT trait estimate θ, whereas in fixed-item tests scale scores are sometimes based on the summed (number-correct) scores. Summed scores might be less sensitive to misfit even when IRT procedures are used for equating the summed scores. Perhaps more significantly, misfitting items early in the CAT might lead to poor choices of later items. The purpose of this research was to examine the effects of misfitting items on θ estimation in CAT.

Sinharay and Haberman (2014) stressed the importance of considering the practical impact of item misfit on the decisions made from test scores. Looking at several testing programs, they examined IRT-based true score equating functions applied to the summed scores. In one example, using two-parameter logistic (2PL) and generalized partial credit models with external anchor items, they compared equating functions using all anchor items versus removing the poorly fitting

items from the anchor set. In another example for three different tests, they compared equating functions based on one-parameter and three-parameter logistic (1PL and 3PL) models, where the 3PL model fit the data better. In a third example for two tests using a 3PL model with internal anchor items, they compared equated scale scores based on the full test versus a test omitting misfitting items. Across all of these examples, differences in the equating functions were relatively small for all or most score points: the magnitude of the differences seemed to be slightly larger for the examples that compared different models than for the examples that compared including/ excluding individual misfitting items. Sinharay and Haberman reported similar results for proportions of examinees consistently classified above/below a cut score: very low rates of different classifications, although slightly higher in the examples that used different models for the pair of equatings.

Another practical context is the estimation of group means. Using several NAEP data sets covering different grade levels and subject areas, van Rijn et al. (2016) computed subgroup means using all items and again omitting the nine worst fitting items. They followed the usual NAEP methods, which involve estimating group means through an IRT regression model. Group means were rounded to the nearest integer (on a 0–500 scale) and the difference between the mean with and without the worst fitting items was never greater than 1.0 for any comparison. They also looked at the percentage of examinees in each score category, by subgroup, and again found no differences greater than 1.0.

Köhler and Hartig (2017) defined practical significance in terms of maximum and minimum correlation with a covariate if misfitting items were omitted. Their procedure can estimate the changes to the correlation without observing the covariate, to aid in planning correlational studies. They noted, "The minimum and maximum change depends on the amount of additional variance that is induced by the misfitting items, on the amount of misfitting items relative to the fitting items, and on the strength of the relationship between the latent variable and the covariate" (p. 391). In their real data example, the potential change in the correlation seemed small.

Overall, these studies showed generally small practical impact of item misfit. But, as Sinharay and Haberman (2014, p. 33) noted, studies of misfitting items in CAT have not been reported. The research questions examined in this study were:

1. When items show misfit near the point of maximum information, how does the proportion of misfitting items impact the bias and standard error of the θ estimates?
2. Does test length interact with the proportion of misfitting items?

The items were chosen to misfit near the point of maximum information because this location of misfit might be particularly problematic in a CAT, especially given that CATs usually have fewer items than fixed-item tests. Items that misfit near an examinee's θ estimate have the greatest potential to cause misestimation. In the later stages of a CAT, examinees should only see items near their θ estimates, so an item that misfit in regions where its information was low would be unlikely to be administered to the examinees who would be most affected by the misfit. As the proportion of misfitting items increased, the accuracy of the θ estimates would generally be expected to decrease. At a given θ, some examinees' θ estimates might be overestimated and others might be underestimated, depending on which misfitting items they received. There might be no systematic bias, but greater random variance than that observed when the items fit the model. For shorter tests, the empirical standard error might be more greatly impacted by increasing misfit than it would be in longer tests because there would be less opportunity to recover from misfitting items early in the test or for errors to cancel across different misfitting items.

Method

Two factors were manipulated: percentage of misfitting items administered (0%, 10%, 30%) and test length (10, 30, 50) for a 3-by-3 design. For each condition, 1,000 simulees were generated at each of nine ability levels: −2.0, −1.5, −1.0, −0.5, 0.0, 0.5, 1.0, 1.5, 2.0.

Test Length and Item Bank Size

The three test lengths (10, 30, and 50 items) were within the ranges used in other studies or in practice. The ASVAB tests each have 10–15 items (Armed Services Vocational Aptitude Battery, n.d.). The GMAT has 31 quantitative and 36 verbal items (Graduate Management Admission Council, n.d.). Formerly, when the GRE was item-level adaptive, it had 28–35 items depending on the section (Mills, 1999). Wyse (2021) described a K-12 reading and math CAT program that had 34 items in each content area. Other simulations have used similar ranges of items: for example, Rudner and Guo (2011) used test lengths of 10, 15, 20, 25, and 30, Han (2009) chose a test length of 40, and Kingsbury and Wise (2020) simulated a test length of 45.

The item bank size was 20 times the test length, yielding 200, 600, and 1,000 items. Gönülateş (2019) noted, "…there are no one-size-fits-all general rules for an adequate-sized item pool" (p. 1136). A classic recommendation is 12–16 times the adaptive test length (Stocking, 1994), based on considerations of content constraints, item exposure, and test overlap. Larger item banks are also needed when examinee θs cover a broader range so that there are adequate items to match the examinees. For example, one low-stakes math item bank had 1,722 items for Grades 2–5 and 1,873 items for Grades 6–12 (Kingsbury & Houser, 1999). A variable-length English test for Chinese university students that averaged 11 items had an item bank of 258 items (He & Min, 2017). In simulations, Şahin and Ozbasi (2017) used 500 items for a variable-length test, and Han (2009) used 500 items for a 40-item test (12.5 times test length). Şahin and Weiss (2015) simulated bank sizes of 100, 200, 300, and 500. Rudner and Guo (2011) crossed test lengths of 10 to 30 with item banks of 200, 300, and 600. Luo and Wang (2019) simulated a bank eight times the test length. Overall, the 20:1 ratio was within the high end of the range used by other researchers.

Items

An item bank was generated for the smallest item bank, and that item bank was duplicated with small perturbations in the parameters to create larger item banks that had nearly identical properties. An item bank of misfitting items was created. Misfit was simulated in two ways. Misfit Type A items had a flat spot in the middle, similar to misfit used in Wells and Bolt (2008) or van Rijn et al.'s (2016) Type 3 items:

$$P_i(\theta) = .2 + .8\frac{\exp(1.5\theta)}{1+\exp(1.5\theta)}\frac{\exp\left[a_i(\theta - b_i)\right]}{1+\exp\left[a_i(\theta - b_i)\right]} + .8\left[1-\frac{\exp(1.5\theta)}{1+\exp(1.5\theta)}\right]\frac{\exp\left[(a_i+0.85)(\theta - b_i+1.5)\right]}{1+\exp\left[(a_i+0.85)(\theta - b_i+1.5)\right]}, \quad (1)$$

where $P_i(\theta)$ is the probability of correct response to Item i, given ability θ, a_i and b_i are the item

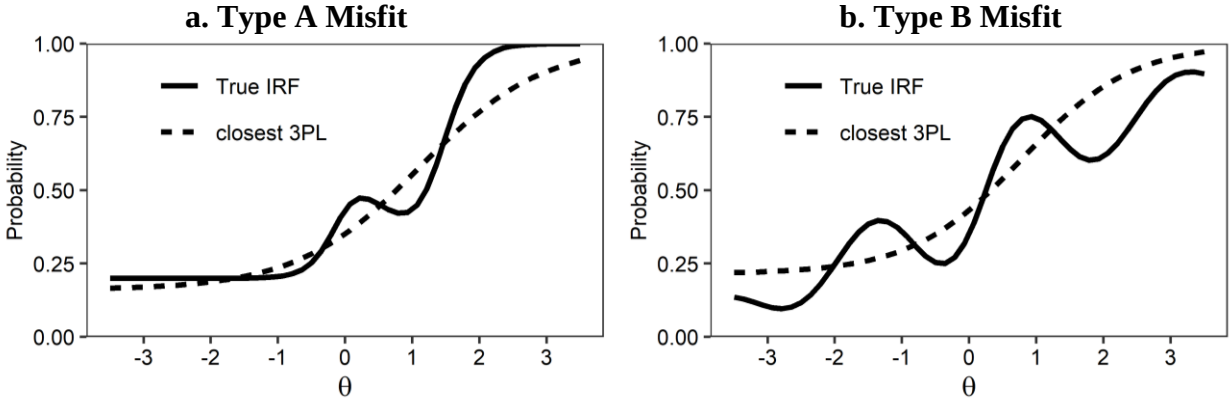
discrimination and difficulty, respectively. An example is shown in Figure 1a. Misfit Type B items (Figure 1b) had a more complex curve:

$$P_i(\theta) = \frac{\exp\left[a_i(\theta - b_i)\right]}{1+\exp\left[a_i(\theta - b_i)\right]} + 0.4\left(0.5 - \left|0.5 - \frac{\exp\left[a_i(\theta - b_i)\right]}{1+\exp\left[a_i(\theta - b_i)\right]}\right|\right)\sin \pi k_i(\theta - b_i), \qquad (2)$$

where $k_i = 0.9 - |b_i|/5$, π is the mathematical constant ≈ 3.14, and other variables are as defined for Equation 1. An example is shown in Figure 1b. The generating parameters for the Type A item were $a_i = 4.252$, $b_i = 1.437$, and the best fitting 3PL parameters were $a_i = 1.093$, $b_i = 1.118$, $c_i = 0.161$, with maximum information at $\theta = 1.33$. The generating parameters for the Type B item were $a = 0.554$, $b = 0.25$, $k = 0.85$, and the best fitting 3PL parameters were $a = 1.219$, $b = 0.791$, $c = 0.215$, with maximum information at $\theta = 1.02$.

Misfit generated by a four-parameter logistic (4PL) model was also considered, but the misfit for 4PL items was large only at θs considerably higher than the item difficulty. Similarly, items where the probability of correct response was at chance level for low θs but dropped below chance for middle θs had large misfit only at θs considerably lower than the item difficulty. In CAT, these items would be unlikely to be administered to examinees in the range of large misfit.

Figure 1. Examples of Misfitting Items



For each misfitting item, the closest 3PL item parameters were also determined. The 3PL function is

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp\left[a_i(\theta - b_i)\right]}{1+\exp\left[a_i(\theta - b_i)\right]}, \qquad (3)$$

where $P_i(\theta)$ is the probability of correct response to Item i, and a_i, b_i, and c_i are the item discrimination, difficulty, and lower asymptote parameters.. The closest 3PL parameters were found by minimizing the distance between the true item response function (IRF) and a 3PL IRF, using the *nlm* function in the *stats* package in base R (Version 4.0.2). These IRFs were termed the true misfitting IRF and the 3PL IRF, respectively. Misfitting items were chosen to cover a wide range

of difficulty and to show misfit near the point of maximum information, where the misfit could potentially be most harmful in a CAT. The point of maximum information, termed here the item *location*, was based on the 3PL IRF. The local misfit was quantified using Wells and Bolt's (2008) misfit index, except integrating only over the region of location ±0.5.[1] This index is the root mean square difference (RMSD) between the 3PL IRF and the true, misfitting IRF:

$$\text{RMSD} = \sqrt{\sum_{\theta_{j=\text{loc}-0.5}}^{\theta_{j=\text{loc}+0.5}} \frac{[P_{\text{fit}}(\theta_j) - P_{\text{misfit}}(\theta_j)]^2}{J}}, \quad (4)$$

where loc is the location of maximum information, $P_{\text{fit}}(\theta_j)$ is the 3PL IRF, $P_{\text{misfit}}(\theta_j)$ is the misfitting IRF, θ_j is one of 11 evenly spaced points in the interval ±0.5 around this location, and J is the total number of points evaluated, set at 11 here.

Values were selected for the item parameters such that the locations spanned the range of −2 to 2, and each item had a high magnitude of misfit near the location of maximum information. Parameters were chosen such that the local misfit (Equation 4) was at least 0.08 (mean = 0.103). After the parameters for the misfitting items were selected for the item bank, corresponding items with the closest 3PL parameters were added to the bank. Thus, for each misfitting item, there was a 3PL item with the same location of maximum information.

Descriptive statistics for the 3PL parameters are displayed in Table 1. In the appendix, Tables A1 and A2, respectively, provide the misfitting and 3PL parameters for each item in the smallest item bank. The *a* parameters seem somewhat lower than would be expected on standardized tests. This was an artifact of creating the misfit near what would be the point of maximum information in a 3PL IRF. This is the point where the slope of the IRF would be steepest, but the misfit tended to flatten the curve. Thus, the closest fitting 3PL IRF tended to have a somewhat low *a* parameter.

Table 1. Descriptive Statistics for Closest Fitting 3PL Items

Parameter	Mean	SD	Minimum	Maximum
a	1.08	0.14	0.92	1.61
b	–0.11	1.07	–2.29	2.04
c	0.08	0.09	0.00	0.33

Simulated Responses, Item Selection, and Scoring

In simulation studies, the item selection and scoring are often based on item parameters re-calibrated from a simulated sample, instead of the true item parameters. This would introduce realistic random error. In this context, the misfitting items could introduce systematic error as well. If there were many items that misfit in the same location and in the same way, they could distort the measurement metric, expanding it in some spots and contracting it in others (Sauder & DeMars,

[1] Because the RMSD was calculated only over a limited range within which the density was fairly uniform, the differences were not weighted by the relative density at each point.

2020, p. 375), thus biasing the item parameter estimates. Any effects would depend on the proportion of misfitting items and the nature of the misfit. Because these effects would likely not generalize, the item parameters were not recalibrated and the true/analytically closest fitting parameters were used for item selection.

The misfitting IRF and the 3PL IRF were used in simulating the misfitting and fitting responses, respectively. In the 0% misfitting condition, responses were generated using the 3PL parameters. In the misfitting conditions, for each simulee 10% or 30% of the item positions were randomly flagged to be misfitting. The item selection and scoring was based on the 3PL model, but the response was generated based on the misfitting IRF. For example, a simulee might be assigned to receive misfitting items in Positions 3, 5, and 10. Responses to Items 1, 2, 4, and 6–9 would be generated from the 3PL IRF, but responses to Items 3, 5, and 10 would be generated from the misfitting IRF. Each simulee's initial θ estimate was set to 0.0. An initial item was randomly selected from among the items with $a_i < 1.7$, and $-0.7 < b_i < 0.7$. To avoid infinite estimated θ, maximum a posteriori probability (MAP) estimation was used, with an $N(0,2^2)$ prior. This relatively diffuse prior was moderately informative for the first few items but had little effect as the test length increased. After the first item, items were selected based on maximum information at the current estimate of θ, with no content constraints or exposure control. After an item was administered, both the misfitting and 3PL versions of the item were ineligible for re-administration to that simulee.

Evaluation

At each of nine fixed θ levels (–2.0, –1.5, –1.0, –0.5, 0.0, 0.5, 1.0, 1.5, 2.0), 1,000 examinees were simulated. For each θ level, within each condition, the mean and standard deviation of the 1,000 estimated θs were calculated. At θ_j, the difference between the true θ and the mean estimate represented bias, and the standard deviation of the estimated θs represented the empirical standard error of θ.

Results

Figure 2 shows the estimated bias of θ. With no misfitting items, for the shortest test (10 items) the most extreme θs were biased inward somewhat due to the prior. This bias was smaller than would be seen on a fixed-item test because the influence of the prior decreases as the information near θ increases. CATs generally have more information for extreme values of θ than do fixed-item tests.

For all test lengths, but especially the 10-item test, there was a small positive bias for $\theta = 0$ in the 0% misfitting condition. This may seem anomalous. In the 0% misfitting condition, there are two things that can cause bias: priors and maximum likelihood estimation (MLE). Because MAP estimation finds the maximum of the posterior likelihood (proportional to the likelihood times the prior), some of the properties of MLE apply. The influence of the priors can cause bias toward the mean of the prior, 0.0. This is likely the source of the bias for other values of θ in the shorter tests, but obviously not for the bias when $\theta = 0.0$. In contrast, MLE tends to bias parameter estimates away from the area with the most information (Lord, 1983). Thus, if a test provides the most information in the low end of the θ range, θ will be overestimated for middle and high θ examinees. Typically in CAT, examinees receive tests well-matched to their θs and there is little bias. In this

study, in the effort to create items with misfit near the location of maximum information, an idiosyncrasy was created in the item bank: The items that had the most information near $\theta = 0.0$ had even more information below 0.0. In Figure 3, for the 10-item 0% misfitting condition, the average information function is shown for the set of items selected at $\theta = 0.0$. The function peaks below $\theta = 0.0$, and, thus, the MLEs are positively biased.

Overall, Figure 2 shows that the absolute value of the bias increased slightly as the proportion of misfitting items increased. Except for at the most extreme values of θ, the absolute value of the difference in bias between 0% and 30% misfitting was < 0.1.

Figure 4 shows the empirical standard errors of the θ estimates (conditional standard deviation of estimates). For the lower values of θ, the standard errors were similar across the different proportions of misfitting items. The small degree of bias induced by the misfitting items was consistent across replications and thus did not increase the conditional standard errors. For the higher values of θ, the standard error showed little or no increase with 10% misfitting items but a slightly larger increase with 30% misfitting items.

Figure 2. Bias in Estimated θ for 0%, 10%, and 30% Misfitting Conditions, as a Function of Test Length

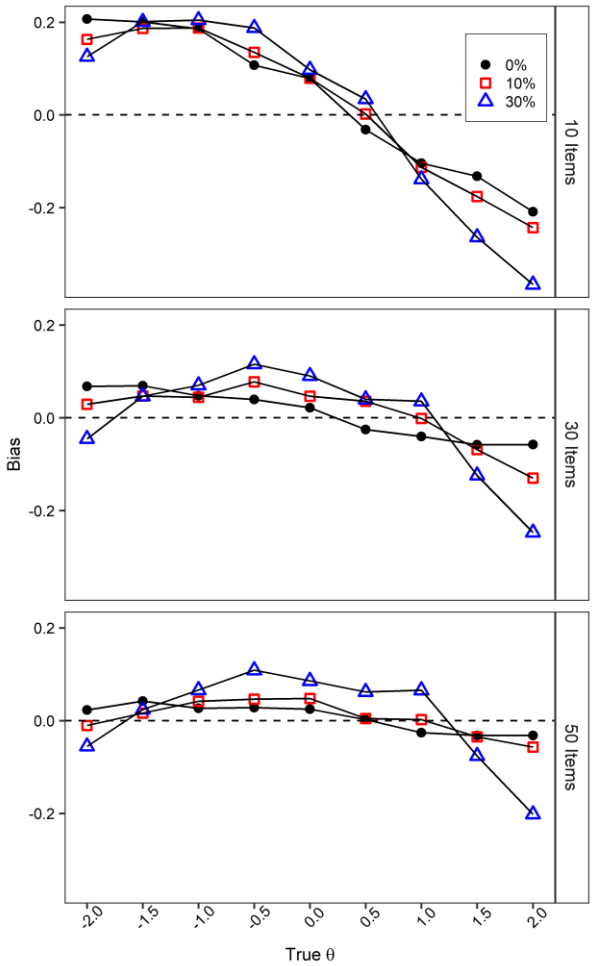


Figure 5 shows the ratio of the estimated standard error to the empirical standard error,

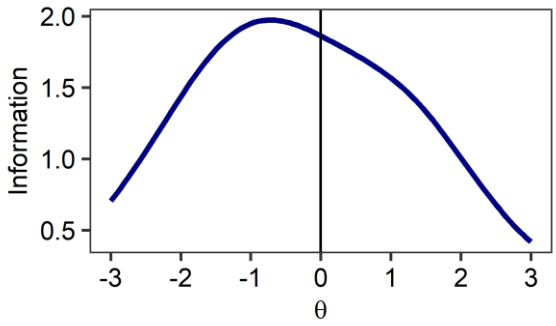
$$\sqrt{\frac{1/I(\hat{\theta})}{\sigma^2(\hat{\theta}\,|\,\theta)}}, \tag{5}$$

where $I(\hat{\theta})$ is the information function for MAP estimates based on the 3PL parameters, evaluated at $\hat{\theta}$,

$$I(\theta) = -E\left(\frac{\partial^2 \ln(L)}{\partial\theta^2} + \frac{\partial^2 \ln[g(\theta)]}{\partial\theta^2}\right), \tag{6}$$

and $\sigma^2(\hat{\theta}\,|\,\theta)$ is the squared empirical standard error of $\hat{\theta}$. In Equation 6, L is the likelihood of the response pattern, $g(\theta)$ is the prior density, and $\partial^2 \ln(L)/\partial\theta^2$ indicates the second derivative of the log-likelihood with respect to θ. The first term is the usual information function for MLEs, and, if $g(\theta)$ is a normal density, the second term is $1/\sigma^2$, where σ^2 is the variance of the prior distribution. A ratio < 1 indicates that the error was estimated to be smaller than it really was, and in a variable-length CAT this would lead to premature termination. For the 30- and 50-item CATs, the ratio was consistently lower for the highest proportion of misfitting items in the upper end of the θ range; the empirical error was large there and the analytically based formulas would underestimate it. For the 10-item tests, the ratio tended to be < 1 for lower values of θ, regardless of whether there were misfitting items.

Figure 3. Average Information Function for Tests Selected for θ = 0, 10 Items, 0% Misfitting



Implications and Conclusions

Administering misfitting items in CAT had little impact on the bias or standard error of the θ estimates. This study was limited by the aspects of the simulation, but most of these aspects were chosen to make it more likely that misfit would have an impact. More realistic choices would be expected to be even more robust to misfit. The fitting items could have higher a_i parameters instead

of matching the slopes of the misfitting items, which were necessarily decreased by the misfit. This would have led to less frequent selection of misfitting items, decreasing the potential for the misfit to impact the θ estimates. Misfit could be simulated at other parts of the IRF, instead of near the item location. This would allow for a greater variety of misfitting functions, but when examinees received misfitting items the misfit would generally have little impact because the misfit would not be near examinees' θs. Content and exposure constraints could make the items less matched to θ; this might increase the impact of misfit further from the information location. But overall, even under conditions specifically chosen to amplify the potential impact of misfit, the effect of misfit on θ estimates was small. This should be reassuring to CAT developers and users.

Figure 4. Empirical Standard Error for θ Estimates by Percentage of Misfitting Items and Test Length

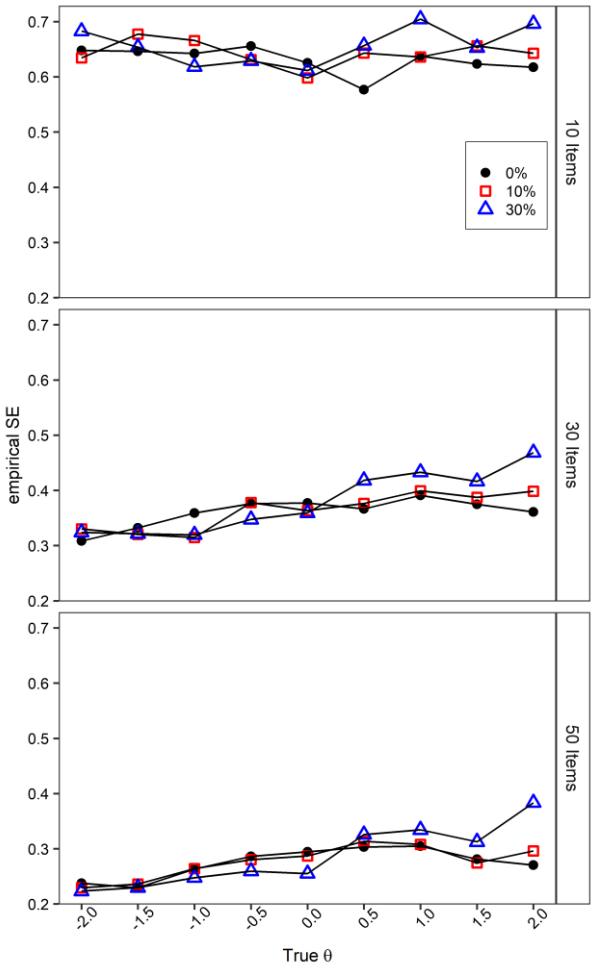
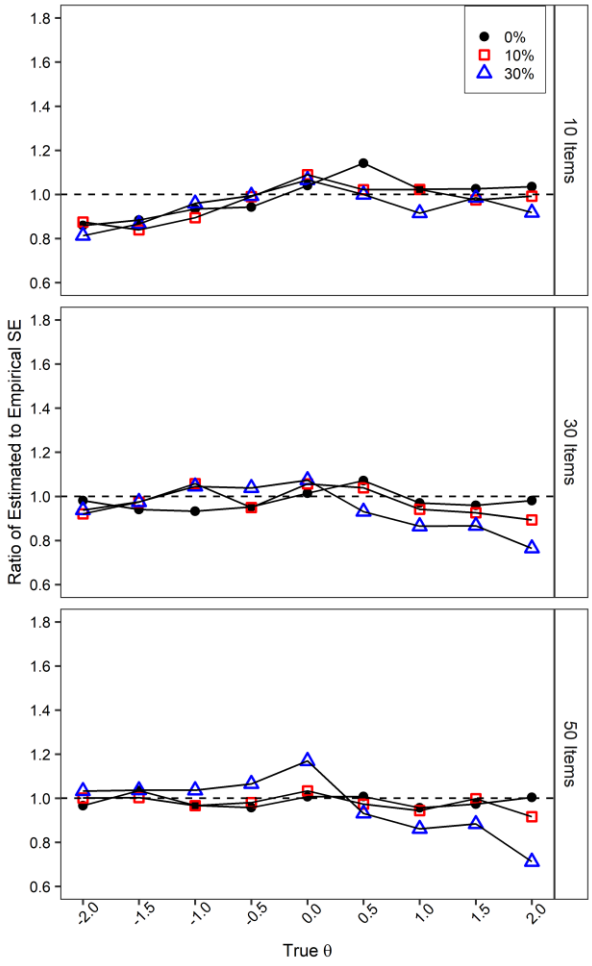


Figure 5. Ratio of Estimated to Empirical Standard Error for θ Estimates by Percentage of Misfitting Items and Test Length



References

Armed Services Vocational Aptitude Battery. (n.d.). *CAT-ASVAB. Applicants*. *WebLink*

Gönülateş, E. (2019). Quality of item pool (QIP) index: A novel approach to evaluating CAT item pool adequacy. *Educational and Psychological Measurement, 79*(6), 1133–1155. *CrossRef*

Graduate Management Admission Council. (n.d.). *GMAT exam: Structure and content. WebLink*

Han, K. T. (2009, June). *Gradual maximum information ratio approach to item selection in computerized adaptive testing* (Research Report No. RR-09-07). GMAC. *WebLink*

He, L., & Min, S. (2017). Development and validation of a computer adaptive EFL test. *Language Assessment Quarterly, 14*(2), 160–176. *CrossRef*

Kingsbury, G. G., & Houser, R. L. (1999). Developing computerized adaptive tests for school children. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 93–115). Lawrence Erlbaum Associates.

Kingsbury, G. G., & Wise, S. L. (2020). Three measures of test adaptation based on optimal test information. *Journal of Computerized Adaptive Testing, 8*(1), 1–19. *CrossRef*

Köhler, C., & Hartig, J. (2017). Practical significance of item misfit in educational assessments. *Applied Psychological Measurement, 41*(5), 388–400. *CrossRef*

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*(2), 233–245. *CrossRef*

Luo, X., & Wang, X. (2019). Dynamic multistage testing: A highly efficient and regulated adaptive testing method. *International Journal of Testing, 19*(3), 227–247. *CrossRef*

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examinations General Test. In F. Drasgow, & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117–135). Lawrence Erlbaum Associates.

Rudner, L. M., & Guo, F. (2011). *Computer adaptive testing for small scale programs and instructional systems* (Research Report No. RR-11-01). GMAC. *WebLink*

Şahin, A., & Ozbasi, D. (2017). Effects of content balancing and item selection method on ability estimation in computerized adaptive tests. *Eurasian Journal of Educational Research, 69,* 21–36. *CrossRef*

Şahin, A., &Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory & Practice, 15*(6), 1585–1595. *CrossRef*

Sauder, D., & DeMars, C. (2020). Applying a multiple comparison control to IRT item-fit testing. *Applied Measurement in Education, 33*(4), 362–377. *CrossRef*

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice, 33*(1), 23–35. *CrossRef*

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (Report No. RR-94-05). ETS. *CrossRef*

van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assessments in Education, 4*(10), 1–23. *CrossRef*

Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. *Applied Measurement in Education, 21*(1), 22–40. *CrossRef*

Wyse, A. E. (2021). How days between tests impacts alternate forms reliability in computerized adaptive tests. *Educational and Psychological Measurement, 81*(4), 644–667. *CrossRef*

Author Address

Christine E. DeMars, Center for Assessment & Research Studies,
James Madison University, 298 Port Republic Road,
MSC 6806, Harrisonburg, VA 22807. U.S.A.
Email: demarsce@jmu.edu

Appendix
Item Parameters for the Smallest Item Bank

Table A1. Type A Items

	Generating Parameters		Closest 3–PL Parameters			Point of Maximum	
Item	*a*	*b*	*a*	*b*	*c*	Information	Local RMSD
1	3.400	–0.400	1.222	–2.261	0.000	–2.261	0.131
2	3.400	–0.360	1.205	–2.211	0.000	–2.211	0.129
3	3.400	–0.320	1.189	–2.160	0.000	–2.160	0.126
4	3.400	–0.280	1.172	–2.108	0.000	–2.108	0.124
5	3.400	–0.240	1.157	–2.055	0.000	–2.055	0.121
6	3.400	–0.200	1.142	–2.002	0.000	–2.002	0.119
7	3.400	–0.160	1.127	–1.948	0.000	–1.948	0.116
8	3.400	–0.120	1.113	–1.893	0.000	–1.893	0.113
9	3.400	–0.080	1.099	–1.837	0.000	–1.837	0.110
10	3.400	–0.040	1.086	–1.781	0.000	–1.781	0.108
11	4.250	0.000	1.136	–1.701	0.000	–1.701	0.120
12	4.250	0.040	1.121	–1.644	0.000	–1.644	0.118
13	4.250	0.080	1.107	–1.586	0.000	–1.586	0.116
14	4.250	0.120	1.093	–1.527	0.000	–1.527	0.114
15	4.250	0.160	1.079	–1.467	0.000	–1.467	0.112
16	4.250	0.200	1.066	–1.406	0.000	–1.406	0.110
17	4.250	0.240	1.053	–1.345	0.000	–1.345	0.108
18	4.250	0.280	1.041	–1.284	0.000	–1.284	0.106
19	4.250	0.320	1.029	–1.221	0.000	–1.221	0.104
20	4.250	0.360	1.018	–1.158	0.000	–1.158	0.102
21	5.100	0.400	1.034	–1.083	0.000	–1.083	0.115
22	5.100	0.440	1.022	–1.019	0.000	–1.019	0.112
23	5.100	0.480	1.011	–0.954	0.000	–0.954	0.110
24	5.100	0.520	1.000	–0.888	0.000	–0.888	0.107
25	5.100	0.560	0.989	–0.822	0.000	–0.822	0.104
26	5.100	0.600	0.978	–0.755	0.000	–0.755	0.100
27	5.100	0.640	0.968	–0.688	0.000	–0.688	0.097
28	5.100	0.680	0.957	–0.621	0.000	–0.621	0.094
29	5.100	0.720	0.947	–0.553	0.000	–0.553	0.091
30	5.100	0.760	0.937	–0.485	0.000	–0.485	0.089
31	5.100	0.773	0.934	–0.462	0.000	–0.462	0.088
32	5.100	0.787	0.931	–0.439	0.000	–0.439	0.087

	Generating Parameters		Closest 3–PL Parameters			Point of Maximum	
Item	*a*	*b*	*a*	*b*	*c*	Information	Local RMSD
33	5.100	0.800	0.927	–0.416	0.000	–0.416	0.086
34	5.100	0.813	0.924	–0.393	0.000	–0.393	0.086
35	5.100	0.827	0.923	–0.364	0.003	–0.358	0.086
36	5.100	0.840	0.925	–0.327	0.008	–0.309	0.086
37	5.100	0.853	0.927	–0.290	0.014	–0.261	0.087
38	5.100	0.867	0.929	–0.253	0.019	–0.213	0.089
39	5.100	0.880	0.932	–0.215	0.025	–0.166	0.091
40	5.100	0.893	0.934	–0.179	0.030	–0.120	0.093
41	5.100	0.907	0.936	–0.143	0.035	–0.075	0.096
42	5.100	0.920	0.939	–0.106	0.040	–0.029	0.098
43	5.100	0.933	0.942	–0.070	0.045	0.015	0.100
44	5.100	0.947	0.944	–0.034	0.050	0.059	0.102
45	5.100	0.960	0.947	0.002	0.055	0.102	0.104
46	5.100	0.973	0.950	0.038	0.059	0.145	0.105
47	5.100	0.987	0.954	0.074	0.064	0.188	0.106
48	5.100	1.000	0.957	0.109	0.068	0.230	0.106
49	5.100	1.013	0.960	0.145	0.073	0.271	0.106
50	5.100	1.027	0.964	0.180	0.077	0.311	0.106
51	5.100	1.040	0.967	0.215	0.081	0.352	0.105
52	5.100	1.053	0.971	0.249	0.085	0.391	0.105
53	5.100	1.067	0.975	0.284	0.089	0.431	0.104
54	5.100	1.080	0.979	0.318	0.093	0.470	0.103
55	5.100	1.093	0.983	0.352	0.096	0.508	0.102
56	5.100	1.107	0.987	0.386	0.100	0.546	0.100
57	5.100	1.120	0.992	0.420	0.103	0.583	0.099
58	5.100	1.133	0.996	0.453	0.107	0.620	0.098
59	5.100	1.147	1.001	0.486	0.110	0.657	0.097
60	5.100	1.160	1.005	0.519	0.114	0.693	0.097
61	5.100	1.173	1.010	0.552	0.117	0.728	0.096
62	5.100	1.187	1.015	0.584	0.120	0.763	0.096
63	5.100	1.200	1.020	0.616	0.123	0.798	0.096
64	5.100	1.213	1.026	0.648	0.126	0.832	0.096
65	5.100	1.227	1.031	0.679	0.129	0.866	0.097
66	5.100	1.240	1.036	0.711	0.131	0.900	0.097
67	5.100	1.253	1.042	0.742	0.134	0.933	0.098
68	5.100	1.267	1.047	0.773	0.137	0.965	0.099
69	5.100	1.280	1.053	0.803	0.139	0.997	0.100

	Generating Parameters		Closest 3–PL Parameters			Point of Maximum	
Item	*a*	*b*	*a*	*b*	*c*	Information	Local RMSD
70	5.100	1.293	1.059	0.833	0.142	1.029	0.102
71	5.100	1.307	1.065	0.863	0.144	1.060	0.103
72	5.100	1.320	1.071	0.893	0.147	1.091	0.104
73	5.100	1.360	1.091	0.980	0.153	1.182	0.109
74	5.100	1.400	1.111	1.064	0.159	1.268	0.114
75	4.250	1.440	1.095	1.124	0.161	1.332	0.098
76	4.250	1.480	1.115	1.204	0.166	1.413	0.101
77	4.250	1.520	1.137	1.281	0.171	1.491	0.105
78	4.250	1.560	1.159	1.356	0.175	1.565	0.109
79	4.250	1.600	1.183	1.429	0.179	1.637	0.112
80	4.250	1.620	1.195	1.464	0.180	1.672	0.114
81	4.250	1.640	1.207	1.499	0.182	1.706	0.115
82	4.250	1.660	1.220	1.533	0.184	1.739	0.117
83	4.250	1.680	1.233	1.567	0.185	1.772	0.118
84	4.250	1.700	1.246	1.600	0.186	1.804	0.120
85	3.400	1.720	1.175	1.601	0.182	1.814	0.094
86	3.400	1.740	1.187	1.634	0.184	1.846	0.096
87	3.400	1.760	1.198	1.667	0.185	1.877	0.097
88	3.400	1.780	1.210	1.698	0.186	1.908	0.099
89	3.400	1.800	1.222	1.730	0.187	1.939	0.100
90	3.400	1.820	1.234	1.761	0.189	1.968	0.101
91	3.400	1.840	1.246	1.791	0.190	1.998	0.102
92	3.400	1.880	1.271	1.851	0.191	2.055	0.105
93	3.400	1.920	1.296	1.909	0.193	2.110	0.107
94	3.400	1.960	1.322	1.965	0.195	2.163	0.109
95	3.400	2.000	1.349	2.020	0.196	2.215	0.110

Table A2. Type B Items

	Generating Parameters		Closest 3–PL Parameters			Point of Maximum	
Item	*a*	*b*	*a*	*b*	*c*	Information	Local RMSD
96	1.190	–1.900	1.315	–2.055	0.000	–2.055	0.095
97	1.178	–1.859	1.303	–2.013	0.000	–2.013	0.095
98	1.166	–1.819	1.291	–1.971	0.000	–1.971	0.095
99	1.154	–1.778	1.280	–1.928	0.000	–1.928	0.095
100	1.142	–1.738	1.269	–1.885	0.000	–1.885	0.095
101	1.130	–1.697	1.259	–1.841	0.000	–1.841	0.095
102	1.118	–1.657	1.249	–1.797	0.000	–1.797	0.095
103	1.106	–1.616	1.239	–1.753	0.000	–1.753	0.094
104	1.094	–1.576	1.229	–1.709	0.000	–1.709	0.094
105	1.082	–1.535	1.220	–1.664	0.000	–1.664	0.093
106	1.070	–1.495	1.212	–1.619	0.000	–1.619	0.093
107	1.058	–1.454	1.203	–1.574	0.000	–1.574	0.092
108	1.046	–1.414	1.195	–1.529	0.000	–1.529	0.092
109	1.034	–1.373	1.187	–1.483	0.000	–1.483	0.091
110	1.022	–1.332	1.180	–1.437	0.000	–1.437	0.091
111	1.010	–1.292	1.172	–1.392	0.000	–1.392	0.090
112	0.998	–1.251	1.165	–1.346	0.000	–1.346	0.090
113	0.986	–1.211	1.158	–1.300	0.000	–1.300	0.089
114	0.974	–1.170	1.151	–1.254	0.000	–1.254	0.089
115	0.962	–1.130	1.144	–1.208	0.000	–1.208	0.089
116	0.950	–1.089	1.137	–1.162	0.000	–1.162	0.088
117	0.938	–1.049	1.130	–1.116	0.000	–1.116	0.088
118	0.926	–1.008	1.123	–1.071	0.000	–1.071	0.088
119	0.914	–0.968	1.116	–1.025	0.000	–1.025	0.088
120	0.902	–0.927	1.109	–0.979	0.000	–0.979	0.088
121	0.890	–0.886	1.102	–0.934	0.000	–0.934	0.088
122	0.878	–0.846	1.094	–0.889	0.000	–0.889	0.089
123	0.866	–0.805	1.086	–0.844	0.000	–0.844	0.089
124	0.854	–0.765	1.078	–0.799	0.000	–0.799	0.089
125	0.842	–0.724	1.069	–0.754	0.000	–0.754	0.090
126	0.830	–0.684	1.059	–0.710	0.000	–0.710	0.091
127	0.818	–0.643	1.050	–0.666	0.000	–0.666	0.091
128	0.806	–0.603	1.039	–0.622	0.000	–0.622	0.092
129	0.794	–0.562	1.028	–0.578	0.000	–0.578	0.093
130	0.782	–0.522	1.016	–0.535	0.000	–0.535	0.094
131	0.770	–0.481	1.003	–0.492	0.000	–0.492	0.095

	Generating Parameters		Closest 3–PL Parameters			Point of Maximum	
Item	*a*	*b*	*a*	*b*	*c*	Information	Local RMSD
132	0.759	–0.441	0.990	–0.449	0.000	–0.449	0.096
133	0.747	–0.400	0.976	–0.407	0.000	–0.407	0.097
134	0.744	–0.390	0.972	–0.396	0.000	–0.396	0.097
135	0.741	–0.380	0.968	–0.386	0.000	–0.386	0.098
136	0.738	–0.370	0.964	–0.375	0.000	–0.375	0.098
137	0.735	–0.360	0.961	–0.365	0.000	–0.365	0.098
138	0.732	–0.350	0.957	–0.355	0.000	–0.355	0.099
139	0.729	–0.340	0.953	–0.344	0.000	–0.344	0.099
140	0.726	–0.330	0.949	–0.334	0.000	–0.334	0.099
141	0.723	–0.320	0.945	–0.324	0.000	–0.324	0.100
142	0.720	–0.310	0.942	–0.310	0.001	–0.308	0.100
143	0.717	–0.300	0.943	–0.288	0.006	–0.275	0.100
144	0.714	–0.290	0.945	–0.266	0.011	–0.243	0.100
145	0.711	–0.280	0.946	–0.245	0.016	–0.213	0.100
146	0.708	–0.270	0.948	–0.223	0.021	–0.182	0.100
147	0.705	–0.260	0.949	–0.201	0.025	–0.151	0.100
148	0.702	–0.250	0.951	–0.179	0.030	–0.121	0.100
149	0.699	–0.240	0.952	–0.158	0.035	–0.092	0.100
150	0.696	–0.230	0.953	–0.136	0.039	–0.062	0.100
151	0.693	–0.220	0.955	–0.115	0.044	–0.033	0.100
152	0.690	–0.210	0.956	–0.093	0.048	–0.005	0.100
153	0.687	–0.200	0.958	–0.072	0.052	0.024	0.100
154	0.684	–0.190	0.959	–0.050	0.057	0.052	0.099
155	0.681	–0.180	0.961	–0.029	0.061	0.079	0.099
156	0.679	–0.170	0.962	–0.008	0.065	0.107	0.099
157	0.676	–0.160	0.964	0.013	0.069	0.134	0.099
158	0.673	–0.150	0.965	0.034	0.073	0.161	0.099
159	0.670	–0.140	0.967	0.055	0.077	0.187	0.099
160	0.667	–0.130	0.968	0.076	0.081	0.214	0.098
161	0.664	–0.120	0.969	0.097	0.085	0.240	0.098
162	0.661	–0.110	0.971	0.118	0.089	0.266	0.098
163	0.658	–0.100	0.972	0.139	0.093	0.292	0.098
164	0.655	–0.090	0.973	0.159	0.097	0.317	0.098
165	0.652	–0.080	0.975	0.180	0.101	0.343	0.098
166	0.649	–0.070	0.976	0.201	0.104	0.368	0.098
167	0.646	–0.060	0.977	0.221	0.108	0.393	0.099
168	0.643	–0.050	0.979	0.242	0.112	0.418	0.099

	Generating Parameters		Closest 3–PL Parameters			Point of Maximum	
Item	*a*	*b*	*a*	*b*	*c*	Information	Local RMSD
169	0.640	–0.040	0.980	0.262	0.115	0.442	0.099
170	0.637	–0.030	0.981	0.282	0.119	0.466	0.099
171	0.634	–0.020	0.982	0.303	0.122	0.491	0.100
172	0.631	–0.010	0.984	0.323	0.125	0.515	0.100
173	0.628	0.000	0.985	0.343	0.129	0.538	0.100
174	0.621	0.023	1.001	0.389	0.138	0.592	0.100
175	0.615	0.046	1.019	0.435	0.147	0.644	0.101
176	0.608	0.069	1.038	0.480	0.156	0.694	0.101
177	0.601	0.092	1.057	0.523	0.164	0.742	0.101
178	0.594	0.115	1.078	0.566	0.172	0.788	0.101
179	0.587	0.138	1.099	0.607	0.180	0.833	0.102
180	0.581	0.162	1.122	0.647	0.188	0.875	0.102
181	0.574	0.185	1.145	0.686	0.195	0.916	0.103
182	0.567	0.208	1.170	0.724	0.203	0.955	0.103
183	0.560	0.231	1.196	0.761	0.210	0.992	0.104
184	0.553	0.254	1.224	0.796	0.216	1.027	0.104
185	0.546	0.277	1.252	0.831	0.223	1.061	0.105
186	0.540	0.300	1.282	0.864	0.229	1.093	0.105
187	0.510	0.300	1.221	0.912	0.236	1.157	0.111
188	0.497	0.323	1.238	0.956	0.244	1.204	0.114
189	0.484	0.346	1.257	0.999	0.252	1.249	0.117
190	0.471	0.369	1.277	1.043	0.259	1.293	0.120
191	0.458	0.392	1.299	1.085	0.267	1.336	0.123
192	0.445	0.415	1.321	1.126	0.274	1.378	0.127
193	0.432	0.438	1.345	1.167	0.282	1.418	0.131
194	0.418	0.462	1.371	1.207	0.289	1.458	0.135
195	0.405	0.485	1.398	1.247	0.296	1.496	0.139
196	0.392	0.508	1.426	1.285	0.303	1.533	0.144
197	0.379	0.531	1.456	1.323	0.310	1.570	0.148
198	0.366	0.554	1.487	1.360	0.316	1.605	0.153
199	0.353	0.577	1.519	1.396	0.323	1.639	0.157
200	0.340	0.600	1.553	1.432	0.329	1.672	0.162