Anchor Set Size for Concurrent Estimation of Item Response Theory Models

Ben & Adam

2025-12-07

Introduction

Item Response Theory (IRT) models are used extensively in the design and scoring of standardized assessments. Each item in an assessment is modeled as having its own item characteristic curve, which expresses the probability that a respondent with a given latent trait value θ will respond correctly (in the case of an assessment with correct/incorrect answers, such as an academic test) or choose a particular scale value (as in the case of a psychological assessment with Likert-scale items). The simplest item response model for dichotomous items is the one-parameter logistic (1PL) or Rasch model, which assumes that items vary only in difficulty, the value of θ at which a respondent has a 50% probability of answering correctly. (This parameter is alternatively described as "location" in some literature, since it describes a specific location on the item characteristic curve.) The two-parameter logistic model (2PL) allows items to additionally vary in discrimination, the slope of the item characteristic curve at its inflection point (de Ayala, 2009). The three-parameter logistic (3PL) model extends the 2PL model by adding a guessing parameter corresponding to the probability that a very low-θ respondent will answer correctly (the y-intercept of the item characteristic curve), which is particularly relevant for multiple-choice items. There is no consensus on the "best" model to use for large-scale assessments – although more complex models may fit real-world data most closely (see, e.g., Robitsch, 2022), ability estimates become dependent on the sample of individuals taking the assessment, which proponents of the Rasch model argue is an undesirable property (Stemler & Naples, 2021). In practice, large-scale assessments vary in the item response models used: the Programme for International Student Assessment (PISA) uses a one-parameter model (Okubo, 2022, p. 11), while the National Assessment of Educational Progress (NAEP) uses two- or three- parameter models depending on question type (NCES, n.d.-a).

In many cases, it is desirable to create multiple versions of a given assessment rather than administering the same set of items to all test-takers. In order to estimate all IRT parameters on the same scale, so that estimated ability scores are comparable across versions, it is necessary for some items to appear on multiple versions of the assessment for calibration. It is possible to first estimate parameters for each assessment version before transforming each set of parameters to a common scale (Haebara, 1980; Stocking & Lord, 1983), or to estimate all parameters simultaneously using Maximum Likelihood techniques. Separate estimation with linking is necessary in many real-world scenarios (for example, when a test is administered yearly, scores from one year must be reported before the next year's test has been administered), but concurrent estimation tends to produce more accurate parameter estimates when it is feasible. Hanson & Béguin (2002) report obtaining parameter estimates with lower RMSE using concurrent estimation in most simulation scenarios tested. Likewise, Kim & Kolen (2007) report lower mean squared error of item characteristic functions under concurrent estimation than the Haebara or Stocking-Lord linking methods across a variety of criterion functions used for linking and underlying ability distributions.

Ideally, assessment designers would choose the lowest number of anchor items needed to obtain accurate parameter estimates, to minimize opportunities for cheating or memorization of answers from a past test version. However, reducing the anchor set too far risks non-identifiability of the IRT model. Some work on anchor set design to date has focused on test linking rather than concurrent calibration (e.g. Vale, 1986; de

Gruijter, 1988; Yang & Houang, 1996), but Wingersky et al. (1987) investigated anchor sets of 10, 20, and 40 items and 85 non-anchor items under concurrent calibration (finding, unsurprisingly, that item parameter estimates based on larger estimates were more stable across simulations). García-Pérez et al. (2010) simulated varying anchor set size and criteria used for selection for a shorter test under the graded-response model, which is used for polytomous items. The current study takes a similar approach to identify appropriate anchor set sizes for dichotomous items under the 2PL model, but differs in that the total number of items per test, rather than the total item bank size, is held constant.

The criteria used for selection of anchor items may also play a role in the efficacy of parameter recovery. Sinharay and Holland (2006a) found that "miditests" – anchor tests with a smaller spread of item difficulties than the overall test – correlated more strongly with overall test performance than standard "minitests" selected to reflect the overall distribution of difficulties, which would lead to more effective parameter estimation. In a follow-up study (Sinharay and Holland, 2006b), they found that miditests resulted in slightly lower bias and RMSE in equating tests across two non-equivalent groups, although the impact of anchor selection strategy was smaller than those of overall test length and anchor size.

Methods

Models were fit using the `mirt` package in R (Chalmers, 2012). `mirt` implements full-information maximum likelihood methods for IRT parameters (Bock & Aiken, 1981; Bock et al., 1988) as an option for model fitting in the presence of missing data. Additionally, the `fscores` function used to produce θ estimates based on a fitted model is able to produce either point estimates (expected a-posteriori, maximum a-posteriori, or maximum likelihood) with standard errors, or a set of multiply imputed plausible values. There is theoretical justification for assessing the accuracy of both point estimates and plausible values – the former are needed to assign scale scores to test-takers based on their responses (which may impact their future educational trajectory in the case of state testing or the SAT/ACT exams), while the latter are frequently part of publicly available datasets used to investigate group differences.

All scenarios tested were based on the 2PL model and used a total test length of 30 items, a total of 1000 test-takers, and four test versions (that is, 250 test-takers per version). For all items, discrimination parameters a were drawn from a lognormal distribution with mean 0 and σ 0.25, and difficulty b was drawn from $\mathcal{N}(0, 1.15)$. These distributions were chosen so that the majority of a values would fall between 0 and 2 and the majority of b values would fall between -3 and 3, noted to be typical values for NAEP (NCES, n.d.-b).

To assess the efficacy of detecting group differences based on plausible values, two conditions were used for test-taker θ distribution: One in which all 1000 individuals had values drawn from $\mathcal{N}(0, 1)$, and one in which half had $\theta \sim \mathcal{N}(0, 1)$ and the other had $\theta \sim \mathcal{N}(0.4, 1)$. In the group difference condition, two of the four test versions were taken by individuals from each of the two distributions. After fitting the IRT model in this condition, ten plausible values were generated for each θ, and used to fit a simple linear regression of group on theta, using the `averageMI` function included with the `mirt` package for pooling results across multiple imputations according to Rubin's rules.

The default `mirt` function used to estimate IRT parameters appears to assume an MCAR model where the group of test-takers assigned to each version are drawn from the same distribution; the `multipleGroup` function is considerably slower but can be used to fit a model in which group ability distributions differ across versions. To test the robustness of `mirt` parameter recovery to violations of the equivalent-groups assumption, simulations were run using both `mirt` and `multipleGroup` for each condition. (We omit the `multipleGroup` results for simulations of no group difference, as these results were very similar to the `multipleGroup` results when a group difference was present.)

Table 1 below summarizes the simulation parameters varied. 100 runs were used for each combination of parameters (5 x 3 x 2 x 2 = 60 total conditions).

Table 1: Simulation conditions

Short Name	Description	Conditions
n_anchor	Number of anchor items (answered by all respondents)	**3, 5, 10, 15, 30** (reference condition – same assessment for all respondents)
anchor_type	Strategy for selecting anchor items from overall item pool	**midi** (items closest to moderate difficulty) **easy** (items with lowest difficulty) **random**
group_diff	Difference in mean theta between test-takers receiving versions 1 and 2 vs. versions 3 and 4	**0, 0.4**
assumption	Assumption of `mirt` function used to estimate parameters	**Equivalent groups** (`mirt` used) **NEAT** (`multipleGroup` used)

Note that for the conditions with `anchor_items=30` there should be no difference in performance across anchor selection strategies because in this condition, all students receive all items. In conditions with smaller anchors, the overall item pool from which anchor items are selected is larger, so the difference in difficulty between anchor items is likely to be more pronounced. This should lead to clearer distinctions in the performance of different anchor selection strategies.

Results

Equivalent Groups

Figures 1-3 below show summary statistics related to parameter recovery for varying anchor set sizes using `mirt` to estimate the 2PL model and obtain theta estimates. For each simulation run, the bias and root mean squared error (RMSE) of estimated values across the 1000 theta parameters, the 30-115 difficulty (a) parameters (more items are needed when the anchor set is smaller, so that all test-takers can receive the same total number of items), and the 30-115 discrimination (b) parameters. Each faintly displayed point indicates the outcome of one of the 100 simulation runs for a condition, while the bolder points (and the numbers above each plot) indicate the average bias/RMSE for the parameter across all 100 runs.

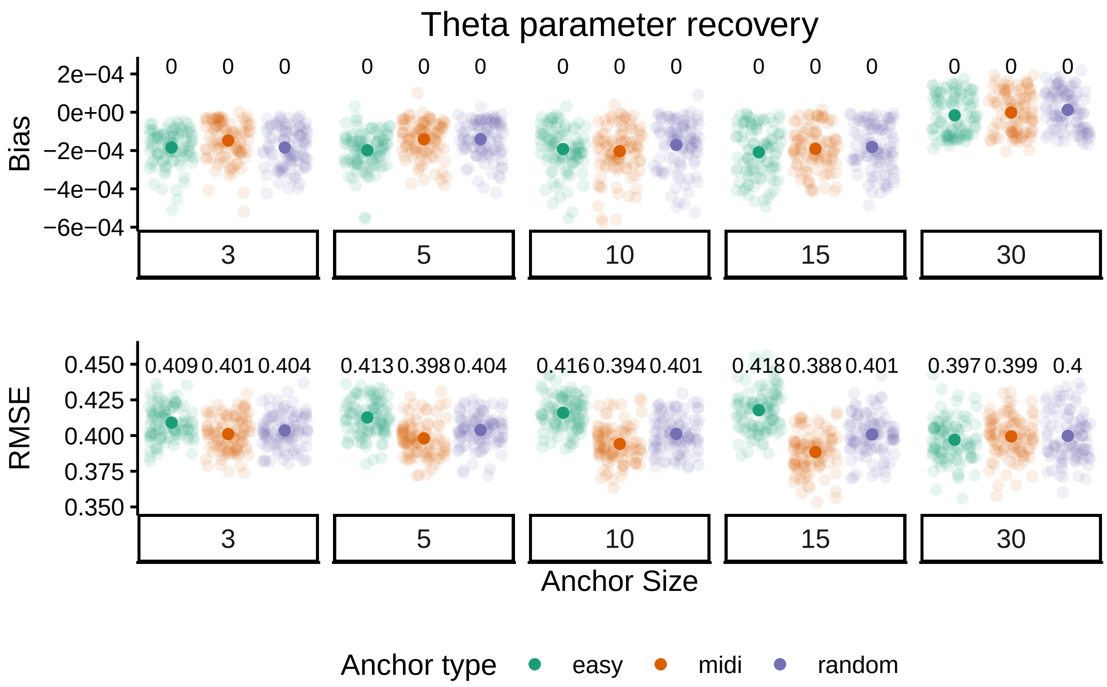


Figure 1: Theta recovery for equivalent groups

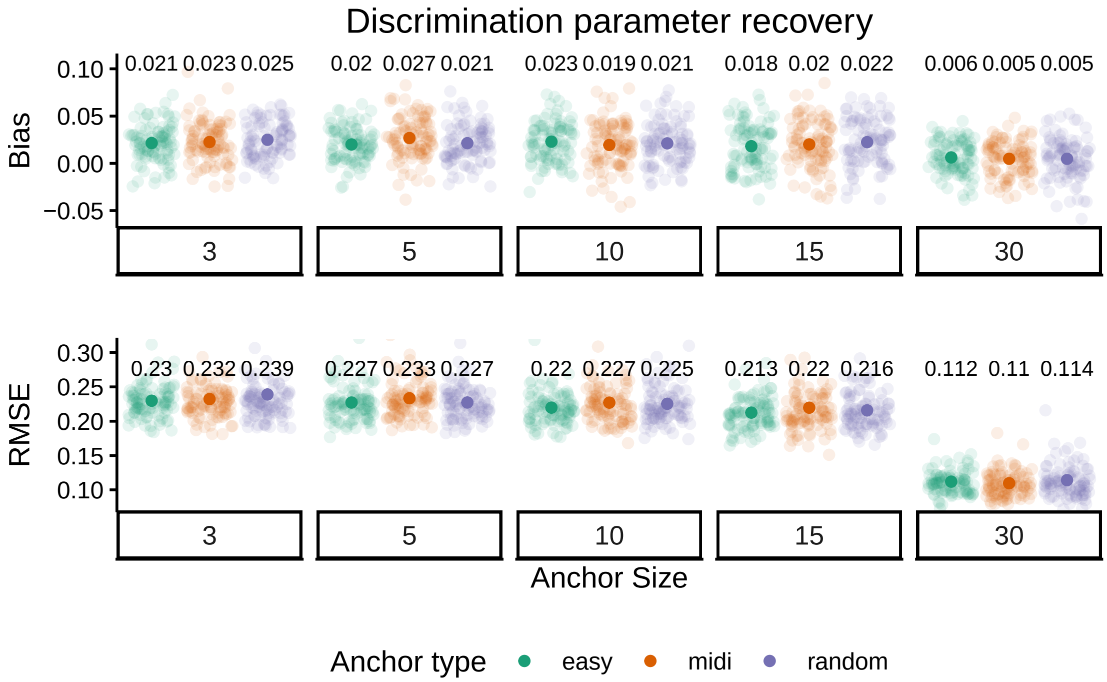


Figure 2: Discrimination parameter recovery for equivalent groups

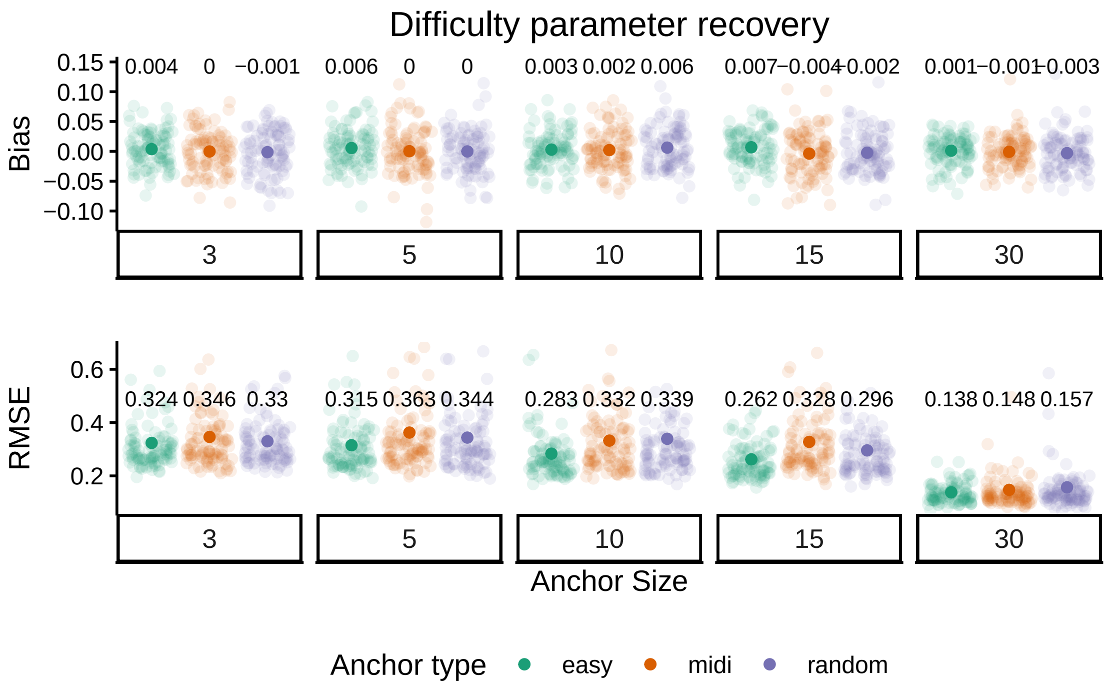


Figure 3: Difficulty parameter recovery for equivalent groups

Surprisingly, there is very little difference in parameter recovery across anchor set sizes. Across conditions, bias in parameter estimates was extremely low.The RMSE for difficulty parameter estimation appears to improve very slightly as the anchor size increases, but this may simply be noise, as the values for the reference (all items shared) conditions show a similar amount of variance across the three anchor selection categories (which have no impact on the outcome in this condition).

Likewise, anchor selection strategies appear to have little, if any, impact on IRT parameter or theta recovery in the equivalent-groups scenario. It appears possible that selecting the easiest items as anchors results in (slightly) more precisely estimated item difficulty parameters but less precise ability estimates, but this again could be noise.

As one might expect based on the parameter recovery, estimated differences between groups based on plausible values were very consistent across conditions:

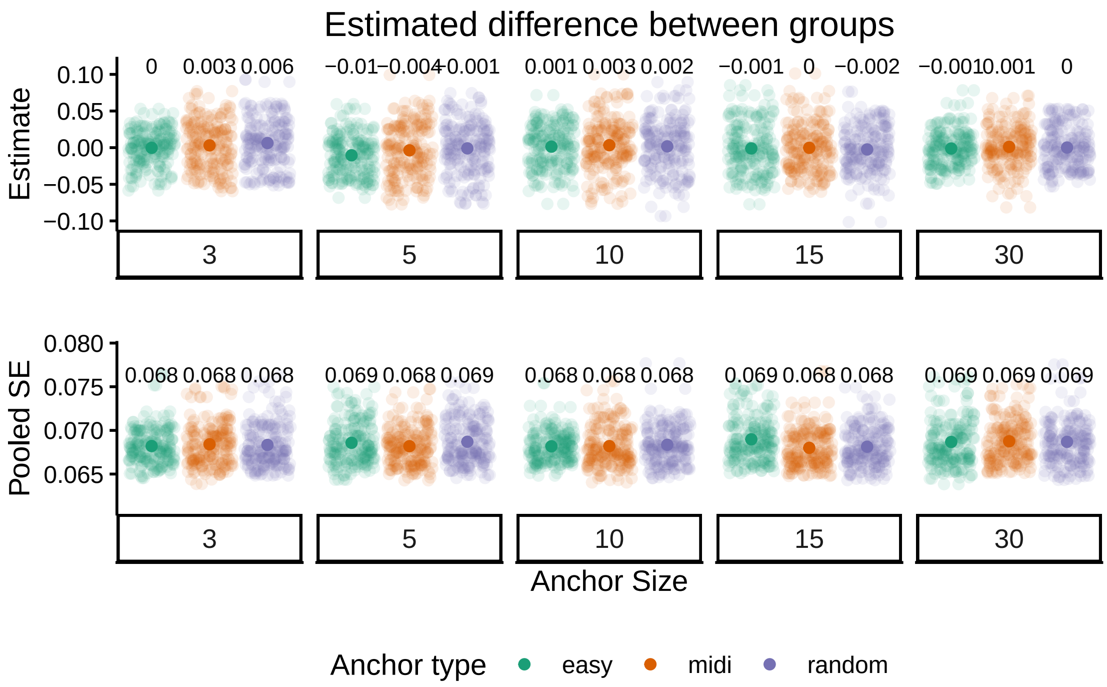


Figure 4: Group difference estimated from plausible values for equivalent groups

Non-equivalent groups under equivalence assumption

Next, we examine the scenario in which test-takers taking different versions are drawn from different ability distributions, but the data are analyzed under the assumption of equivalent groups (using `mirt` rather than `multipleGroup`). Ideally analysts would be aware that groups were non-equivalent and use the appropriate method, but it is plausible and worth investigating whether different anchoring strategies vary in their robustness to model misspecification, knowing that they do not appear to vary when the model is correctly specified.

Estimation of difficulty and discrimination parameters remained accurate in this case, but ability parameters were consistently biased downwards (since the `fscores` function in `mirt` estimates ability parameters centered at 0, whereas the average across two groups with means of 0 and 0.4 is 0.2). The variance in estimated θ was consistent across anchor lengths and selection criteria, as in the case where the equivalent-groups assumption was met.

Where anchor length *did* have an impact, however, was in the group difference estimated after generating plausible θ values. Specifically, while the difference was somewhat underestimated even in the reference condition of all respondents answering the same 30 questions, smaller anchor sets resulted in notably lower estimates. Further, when anchor sets were small, selecting moderate-difficulty items as a “midi” anchor set considerably outperformed selecting the easiest items, and slightly outperformed random selection of anchor items. As expected, these differences were less noticeable when more items were included (and thus higher variability in difficulties was present) in the anchor set.

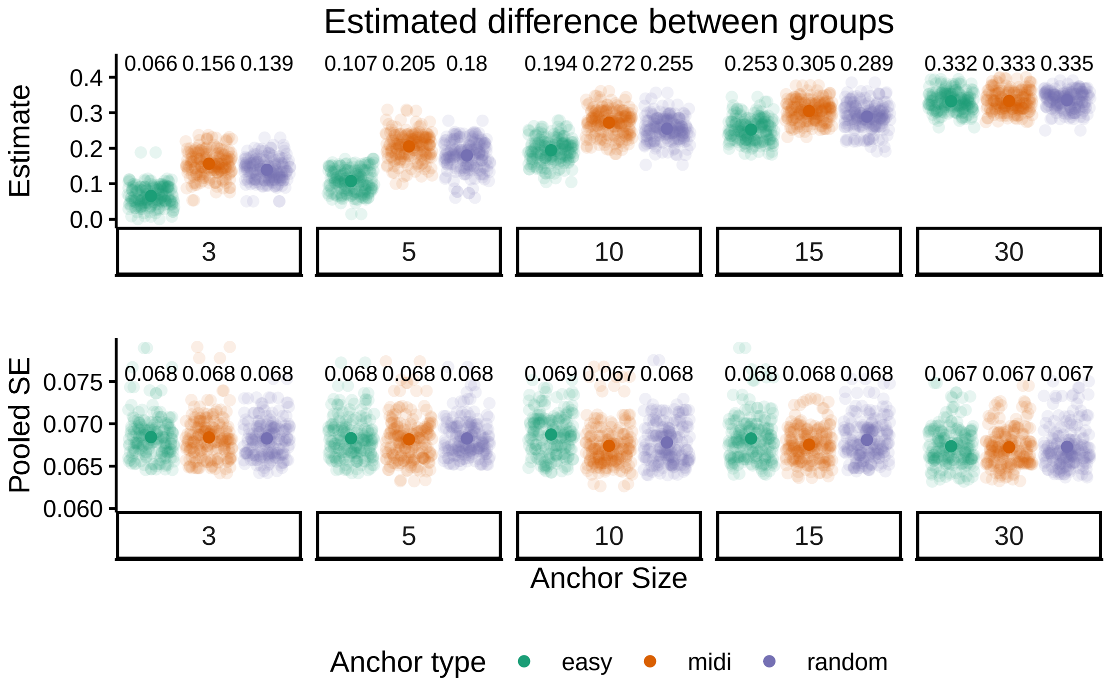


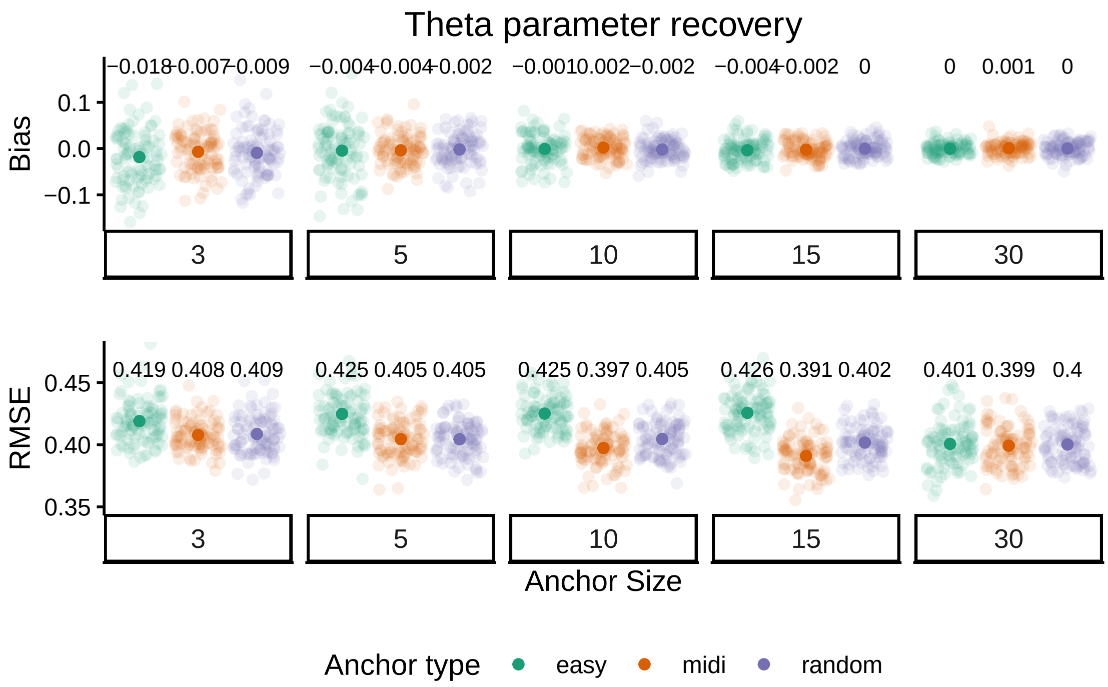
Figure 5: Estimated group differences based on plausible values when group equivalence assumption is violated

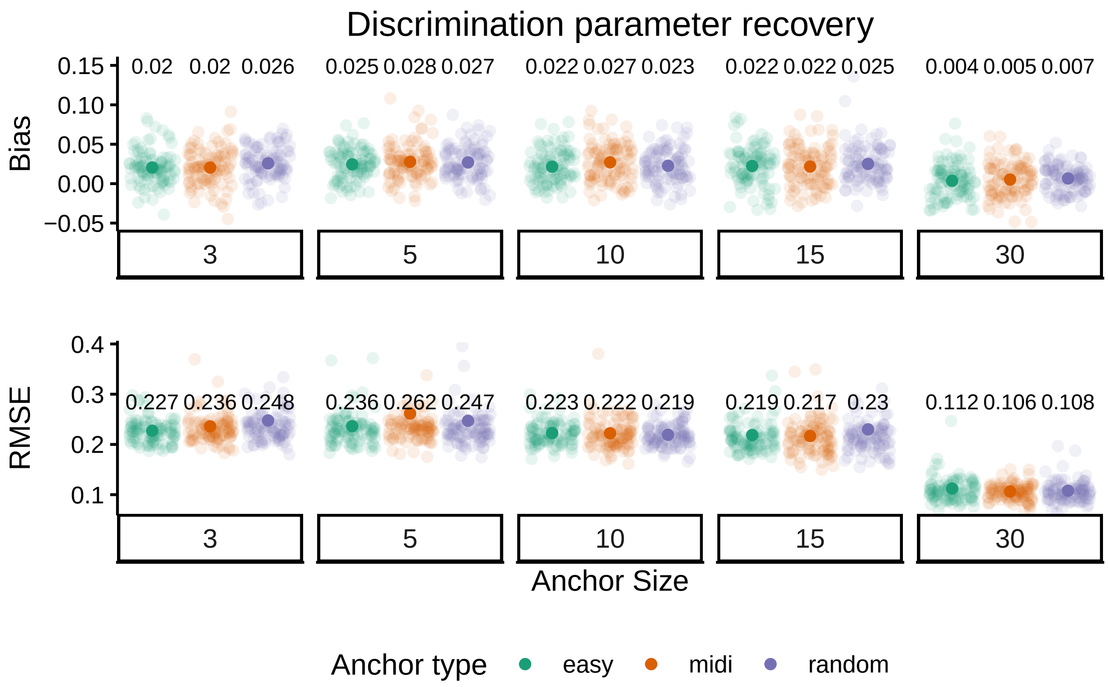
Non-equivalent groups

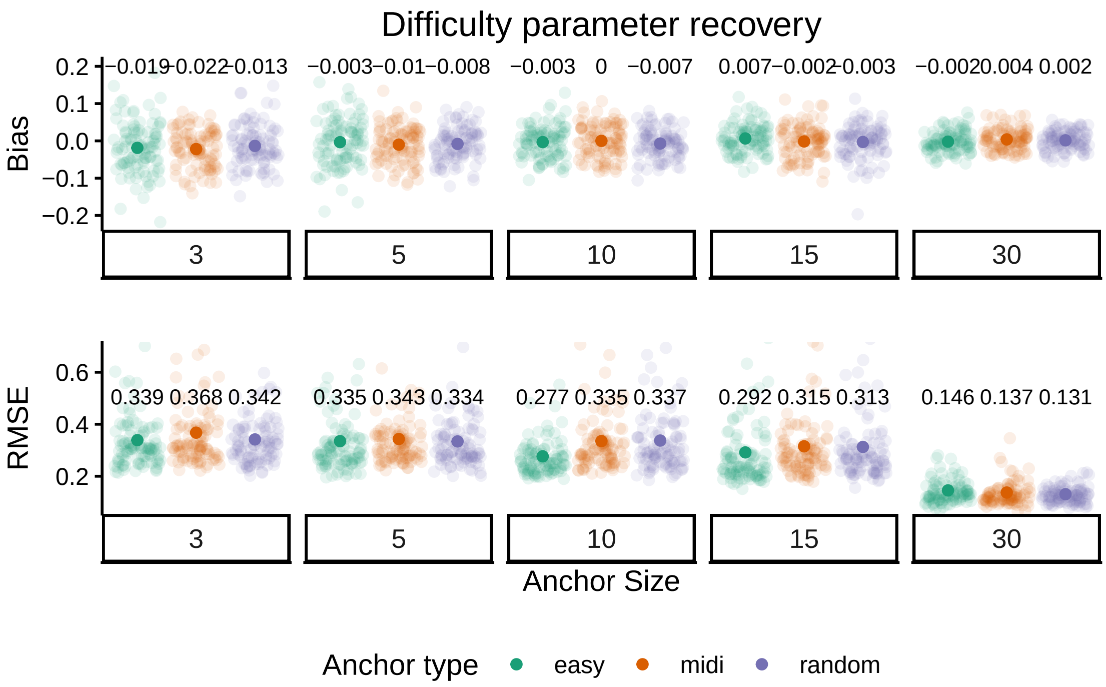
Finally, we examine the case in which groups are non-equivalent and the difference is accounted for in analysis through the `multipleGroup` function in `mirt`. The results presented below are based on completed simulation runs, but it is important to note that the 3-anchor-item condition frequently resulted in models that failed to converge when the easiest items were chosen as anchors. This is due to the fact that when a large item pool is available, the three easiest items are likely to be answered correctly by nearly all test-takers, and this lack of variability in the anchor responses leads to a non-identified IRT model. Simulation runs for this condition were discarded 186 times due to non-convergence or clearly problematic estimates (item parameters estimated to be many times larger than plausible, such as difficulty or discrimination higher than 10).

Larger anchor sets and alternate anchor selection strategies also occasionally resulted in model non-convergence, but this did not occur more than 10 times for any other conditions, including where anchor selection used the 3 easiest items but `mirt` was used to estimate model parameters (which led to only 4 failed runs when no group difference was present and 3 when a group difference was present). This may be because the equivalent groups assumption in `mirt` means there is no need for equating between different test versions. Indeed, `mirt` will return parameter estimates even when provided a set of responses with *no* overlap between test versions, which should result in an unidentified imputation model. Presumably, the item parameter estimates in this case are based only on the responses from test-takers who responded to the item.

The simulation results from 100 runs per condition where the estimation model converged are presented below.


Theta parameter recovery
Bias
0.1
0.0
−0.1
−0.018
−0.007
−0.009
−0.004
−0.004
−0.002
−0.001
0.002
−0.002
−0.004
−0.002
0
0
0.001
0
3
5
10
15
30
RMSE
0.45
0.40
0.35
0.419
0.408
0.409
0.425
0.405
0.405
0.425
0.397
0.405
0.426
0.391
0.402
0.401
0.399
0.4
3
5
10
15
30
Anchor Size
Anchor type
easy
midi
random


Discrimination parameter recovery
Bias
0.15
0.10
0.05
0.00
−0.05
0.02
0.02
0.026
0.025
0.028
0.027
0.022
0.027
0.023
0.022
0.022
0.025
0.004
0.005
0.007
3
5
10
15
30
RMSE
0.4
0.3
0.2
0.1
0.227
0.236
0.248
0.236
0.262
0.247
0.223
0.222
0.219
0.219
0.217
0.23
0.112
0.106
0.108
3
5
10
15
30
Anchor Size
Anchor type
easy
midi
random



Estimates of the item parameters themselves remained minimally impacted by anchor selection. While there was very little difference in the quality of θ estimation when averaging across all 100 runs, individual runs showed more variability with smaller anchor sets. With small anchors it was not uncommon to see runs where the bias of θ estimates was around 0.1. This variability in θ estimation translated into variability in estimates of group difference: while the regression model returned consistent pooled SEs across conditions, the variance of estimated group differences across runs was higher for smaller anchor conditions, and for easy-items anchors compared to medium-difficulty or random selection.

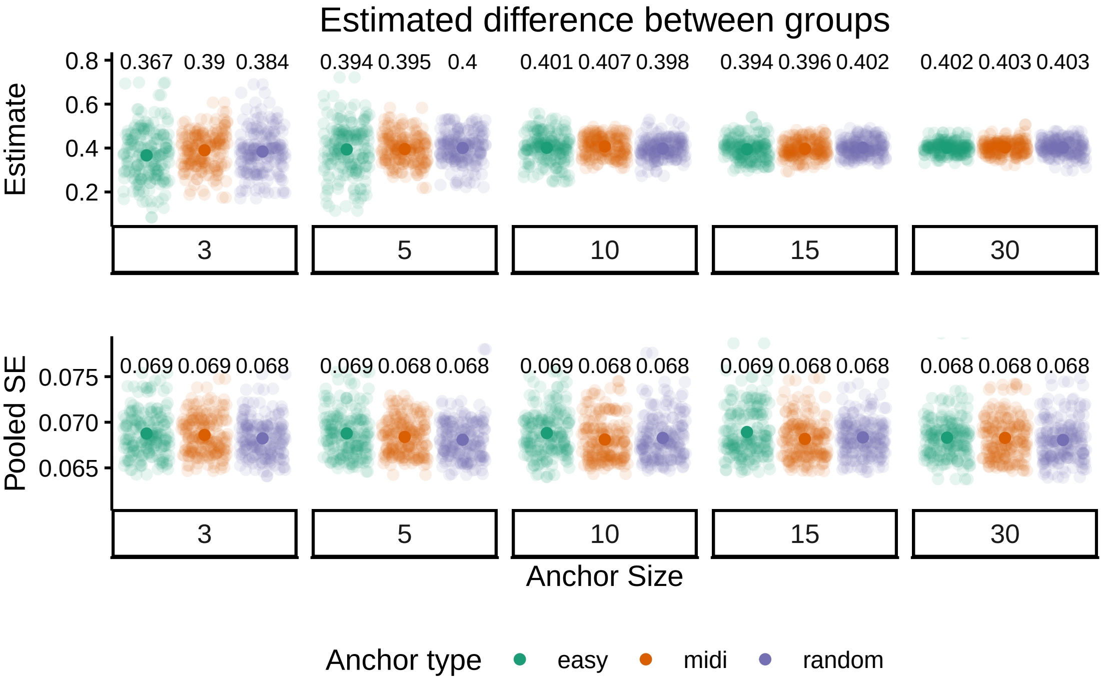


Figure 6: Estimated group differences based on plausible values under non-equivalent groups assumption

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. https://doi.org/10.1007/BF02293801

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261-280. https://doi.org/10.1177/014662168801200305

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford Press.

de Gruijter, D. N. M. (1988). Standard errors of item parameter estimates in incomplete designs. *Applied Psychological Measurement, 12*(2), 109–116. https://doi.org/10.1177/014662168801200201

García-Pérez, M. A., Alcalà-Quintana, R., & García-Cueto, E. (2010). A Comparison of Anchor-Item Designs for the Concurrent Calibration of Large Banks of Likert-Type Items. *Applied Psychological Measurement, 34*(8), 580–599. https://doi.org/10.1177/0146621609351259

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149.

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*(1), 3–24. https://doi.org/10.1177/0146621602026001001

Kim, S., & Kolen, M. J. (2007). Effects on Scale Linking of Different Definitions of Criterion Functions for the IRT Characteristic Curve Methods. *Journal of Educational and Behavioral Statistics, 32*(4), 371-397. https://doi.org/10.3102/1076998607302632

National Center for Education Statistics. (n.d.-a). NAEP Technical Documentation: Item Scaling Models. Retrieved November 23, 2025 from https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_models.aspx

National Center for Education Statistics. (n.d.-b). *NAEP Technical Documentation: NAEP Assessment IRT Parameters.* Retrieved November 23, 2025 from https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_irt.aspx

Okubo, T. (2022). *Theoretical considerations on scaling methodology in PISA.* OECD Education Working Papers, no. 282. https://doi.org/10.1787/c224dbeb-en

Robitzsch, A. (2022). On the choice of the item response model for scaling PISA data: Model selection based on information criteria and quantifying model uncertainty. *Entropy, 24*(6), 760. https://doi.org/10.3390/e24060760

Sinharay, S., & Holland, P. (2006). The correlation between the scores of a test and an anchor test. *ETS Research Report Series, 2006*(1). https://doi.org/10.1002/j.2333-8504.2006.tb02010.x

Sinharay, S., & Holland, P. (2006b). Choice of anchor test in equating. *ETS Research Report Series, 2006*(2). https://doi.org/10.1002/j.2333-8504.2006.tb02040.x

Stemler, S. E. & Naples, A. (2021). Rasch measurement v. Item Response Theory: Knowing when to cross the line. *Practical Assessment, Research, and Evaluation, 26*(11). https://doi.org/10.7275/v2gd-4441

Stocking, M. L., & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement, 7*(2), 201-210. https://doi.org/10.1177/014662168300700208

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10*(4), 333–344. https://doi.org/10.1177/014662168601000402

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). Specifying the characteristics of linking items used for item response theory calibration. *ETS Research Report 1987*(1). https://doi.org/10.1002/j.2330-8516.1987.tb00228.x

Yang, W.-L., & Houang, R. T. (1996). *The Effect of Anchor Length and Equating Method on the Accuracy of Test Equating: Comparisons of Linear and IRT-Based Equating Using an Anchor-Item Design.* Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996). https://eric.ed.gov/?id=ED401308