Not so NEAT? Evidence of Fraudulent Preparation on Internal Anchor Items

N. Scott Bishop

ACT, Inc.

Paper presented at the 2013 Conference on Statistical Detection of Potential Test Fraud in Madison, WI.

DRAFT VERSION. Please do not cite without permission from scott.bishop@act.org. This work is the author's own and does not necessarily represent the thoughts or opinions of ACT, Inc.

Abstract

When items from prior tests are reused and count toward student scores (as can occur in some equating designs), students whose teachers utilize those prior test items to prepare students for future tests can be advantaged. Because such students would generally perform better over the linking items than a well-fitting latent scaling model would predict, the difference between their observed and expected item performances would result in positive residuals. This paper uses Rasch residuals compared across anchor and nonanchor items to investigate how pervasive this type of unethical test preparation may be for one state testing program. Although students in many schools performed better on anchor items, a similar number of schools performed better on nonanchor items. Positive Rasch residuals in the latter case may have resulted from several possible causes (e.g., inappropriate erasures by school staff, prior exposure of nonanchor items when they were embedded field-test items). Additional investigation is required whenever one employs residual analysis in order to determine the exact cause of any aberrant outcomes.

Not so NEAT? Evidence of Fraudulent Preparation on Internal Anchor Items

The integrity of test results has received considerable attention over the last several years. There have been symposiums at NCME's annual conference, a *Request for Information* and follow-up meeting with the U.S. Department of Education, and special conferences on test fraud. NCME has even taken the infrequent step of preparing an organizational statement on this matter (see Bishop et al., 2012). There is very good reason for all this attention. When egregious misconduct occurs on the part of schools and/or educators, valid interpretations regarding educational policy effectiveness simply cannot be maintained. As we move into the next generation of assessments, all stakeholders will have a vested interest in ensuring that improper educator behavior does not dilute the value of test results.

Although there have been many methodological studies over the last several years, they have frequently concerned erasure analysis and copying indices. Although these are widely applied procedures, they are not sensitive to all types of misconduct. Additionally, because of increased media coverage on erasures and wider application of erasure analysis, this particular type of artifice may be on the decline. Still, educators who are motivated to cheat the system will almost certainly do so, and they will likely become more opportunistic and sophisticated in how they do it. Although there are many ways that educators might try to cheat, this paper focuses on one that appears especially risky.

In large-scale testing programs, operational scores are equated to ensure that the scores are comparable across years. To equate scores, a large number of testing programs use a nonequivalent-groups anchor test (NEAT) design that repeats a subset of items across two adjoining years. That is, some items that appeared on the previous year's test also appear on the current year's test. Significantly for this paper, these items often contribute to the students' operational scores in the second administration. In the jargon of test score equating, these items are referred to as *internal* anchor items.

Of course, an obvious danger with such a design is that some unscrupulous school personnel might copy or reproduce items from the old test and then use that information in their instruction, test preparation, and/or coaching activities before future testing. Clearly, students who have educators who engage in these actions will have an unfair advantage on the internal anchor items than other students and, thus, will earn higher test scores.

Literature Review

The mantra from the literature is that any test preparation method that attempts to increase student test scores without increasing student knowledge in the larger domain raises ethical questions. The negative consequence of doing so is to limit the range of inferences one can make to a larger domain based on test performance. Specifically, Mehrens and Kaminski (1991) declare:

> If one wishes to infer to a broader domain from a sample of items (or objectives), then teaching directly to the items or the specific sample of objectives lowers the validity of the inference (p. 14).

Survey results indicate there is good reason to be concerned about this issue. In a survey of Iowa teachers, Lai and Waltman (2008) found that some educators reported using items from prior test forms in their test preparation (median for elementary school teachers = 11%, median for middle school teachers = 16%, median for high school teachers = 17%). Results from the *Survey of State TILSA Member on Test Security* (reported in Olsen and Fremer, 2013) indicate that 63.6% of the surveyed members were concerned about lost or stolen test booklets. Additionally, 90.9% of surveyed members were concerned about teachers coaching students on specific questions before testing. These are especially suggestive results, considering the TILSA survey's response rate was 91.7%.

Olsen and Fremer (2013) discuss the use of person-fit analysis as a data forensic technique. They note this type of analysis would be sensitive to teachers using test questions to prepare students. Across-year performance on internal equating items is routinely compared at the state level. Specifically, vendors commonly screen for "unstable" anchor items during test score equating. But analyses to detect suspicious variations at the class and school levels are generally not done (or at least not made public). As seen below, it is a fairly simple matter to do so using Rasch residuals aggregated at the group level.

Methods

Data

Four years of student-level data containing vectors of item responses were obtained by special permission for this study. (The source of the data will be kept strictly confidential; however, the data is from a large-scale state assessment program whose results were used for NCLB.) As noted in Table 1, the data file contains basic test information (subject, grade, administration year) and unique indicators for districts and schools so that results could be disaggregated at these levels. (Teacher/class names were not provided for obvious reasons.) Supplied documentation indicated which items were unique operational items and which were repeated, anchor items. Data for math and reading were available at seven grade levels while science data was available at three grade levels.

Table 2 provides information about the number of total items and the percentage of those items which were internal anchor items. It should be noted that the tabled information is only approximate to further protect the identity of the testing program. For simplicity, only multiple-choice items were used in this study. In all, the item counts and proportions of internal anchor items are representative of what may be seen in practice.

Analysis

If a teacher used items from the previous year's test to prepare students for upcoming tests, analysis of the item response residuals (the difference between observed and expected item performance) should detect this. For this study, the expected performances for the items were derived with the Rasch model[1]. From Wright and Stone (1979), the probability of a correct response for a dichotomous MC item is:

$$P\{X_{vi}=1\}=\frac{e^{\beta_v-\delta_i}}{1+e^{\beta_v-\delta_i}}$$

The discrepancy, or Rasch model residual, between the expected ($P\{X_{vi} = 1\}$) and actual (X_{vi}) performance for person v on item i is: $Y_{vi} = X_{vi} - P\{X_{vi} = 1\}$. While at the individual response level we might only be surprised by a very able student missing a very easy item or vice versa (Smith, 2000), trends can become more relevant by aggregating results over sets of students and/or items. By aggregating these residuals over all anchor items for individual schools, an index suitable for flagging classes or schools that do appreciably better over the anchor items than the model suggests might result. Although this method alone might be appropriate for flagging suspicious cases when anchor residual averages are compared to those of other classes or schools, the results might be bolstered by comparing the aggregated residuals over anchor items to aggregated residuals over unique (non-equating) operational items and/or field-test items.

[1] It should be noted that expected item performance could be derived from other latent scaling models. The Rasch model was selected in this case for its parsimony. The R package eRm (Mair, Hatzinger, and Maier, 2012) was used to conduct all Rasch analyses for this paper. For each individual test, all items were calibrated together (both anchors and nonanchors).

Results

Tables 3, 4 and 5 provide descriptive statistics for the difference in average residual between anchor and nonanchor items over administration year, grade level, and subject. The nonanchor residual average for each school was subtracted from its corresponding anchor residual average. Therefore, if the average anchor residual for a school was larger than the average residual for nonanchor items, the result would be positive. (Schools with less than 10 students were omitted from the analysis.) Inspection of the maximum and minimum average residual differences indicates that extreme differences developed because large positive residuals occurred for both anchor items and nonanchor items.

Table 6 provides the number of schools that had an average residual difference greater than three standard-deviation units (in absolute value) on at least three occasions over all subjects, grades and years. Differences using estimated thetas area also provided. Such analyses might help identify "frequent offenders" in terms of those schools having many aberrant results over all administered tests. Again, there were just as many schools that had average residual differences that were less than three SD units (indicating higher performance on nonanchor items) as schools that had positive average residual differences greater than three SD units.

Figures 1 – 3 provide scatterplots of each school's residual average for anchor and nonanchor items plotted on the difference in the average theta estimate between anchor and nonanchor items. All differences were computed subtracting the nonanchor result from the anchor result. Over all grades, years, and subjects, the correlation between the residual and theta difference was 0.84. The bivariate plots further confirm that extreme results occur when anchor items have large positive residuals and when nonanchor items have large positive residuals.

Figures 4, 5, and 6 show the mean residual difference for each item for a selected test from each subject area. For the three selected tests, the most extreme result favoring anchor and nonanchor items are presented. These results further reinforce the prior observation that large positive residuals occur over both anchor and nonanchor items.

Discussion

Genuine cases of cheating may go unnoticed because serious, broad-based efforts are needed to detect them. To date, many testing programs may have foregone undertaking all of needed analyses. The aforementioned TILSA survey indicated only 27.3 of respondents' states consistently conducted statistical/psychometric analyses of test responses to detect indications of test security concerns.

Perhaps fit analyses, like response residuals, should be routinely used by testing programs to detect potential cheating (just like erasure and copying analyses currently are). This study was intended as a "proof of concept" of such a procedure for investigating response residuals for internal anchor items. Results indicated that there were just as many schools with large positive residuals over nonanchor items as there were schools with large positive residuals for anchor items. Many factors might cause positive residuals in addition to the treat that exposed anchor items present. These include inappropriate actions like school staff erasures and student copying[2].

As with all flagging procedures, additional analyses are needed to clarify findings and rule out the occurrence of such issues. For example, charts like Figures 4 – 6 can be prepared that indicate the proportion of wrong-to-right erasures for each test item. If peaks occur in both the residual and erasure charts for the same items, that might suggest the positive residuals are due to staff erasures. (Erasure information was not requested from the state that provided data for this study, so such an analysis could not be conducted).

Finally, programs that embed field-test items also risk exposing those items. The testing program that provided data for this study does embed field-test items. Comparing Figures 4 – 6 against information regarding the date of each item's last exposure would be informative. (This information was not requested for this study). It is suggested that anyone using fit analyses to detect threats to test data integrity be prepared to conduct such follow-up investigations. If anchor items are found to consistently yield positive residuals, different equating designs could be considered. Using external anchors that do not contribute to student scores would help some. The equating adjustment might be slightly biased in such a case, but that would impact all schools uniformly. Operational scores would not be boosted on a school-by-school basis. Exposure of field-test items could be mitigated as well. Using separate stand-alone events to try out new items is one option. Embedding might still be useful if forms containing these items

[2] Investigation of model fit should be conducted as well. Technical documents for this testing program indicated that the Rasch model fit the data well.

were assigned to districts versus being spiraled to students in individual classrooms. Although some items could be exposed this way, it would be much more limited within individual schools relative to the exposure of all field-test items through a spiraling process within classrooms.

References

Bishop, N. S., Huff, K., Mitchell, K., Rose-Bond, S., Stemmer, P., Trent, E. R., & Wollack, J. (October, 2012). *Testing and data integrity in the administration of statewide student assessment programs*. National Council on Measurement in Education.

Lai, E. R., & Waltman, K. (2008). Test Preparation: Examining Teacher Perceptions and Practices. *Educational Measurement: Issues and Practice*, *27*(2), 28-45.

Mair, P., Hatzinger, R., and Maier¸M.J. (2012). *eRm: Extended Rasch Modeling*. Version 0.15-1.

Mehrens, W., & Kaminski, J. (1989). Methods for improving standardized test scores: fruitful, fruitless, or fraudulent? *Educational Measurement: Issues & Practice*, *8*(1), 14-22.

Olson, J., and Fremer, J. (2013). *TILSA Test Security Guidebook: Preventing, Detecting, and Investigating Test Security Irregularities*. Washington DC: Council of Chief State School Officers.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, *1*(2), 199-218.

Table 1.

Description of the Data Frame

Study Factors/Variables	Description
Test Subject	M=Math; R=Reading; S=Science.
Grade Level	3, 4, 5, 6, 7, 8, and 11
Administration Year	Year 1, Year 2, Year 3, Year 4
District Indicator	Unique District Indicator
School Indicator	Unique School Indicator
Vector of Item Responses	0, 1 MC, score point for unique items and anchor items

Table 2.

Approximate Number of Items and Percentage of Internal Anchor Items

Subject	Total MC Items	Anchor Percent
Subject Area 1	~ 60	27 to 34
Subject Area 2	~ 40	28 to 42
Subject Area 3	~ 55	28 to 34

Note. Exact item counts not tabled to protect identity of source data.

Table 3.

Distribution of Difference in Average Person Residual (Anchor – Unique) by Year

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Year One					
-0.1321	-0.0144	0.0002	-0.0001	0.0145	0.1388
Year Two					
-0.1565	-0.0129	0.0006	0.0001	0.0136	0.1502
Year Three					
-0.1557	-0.0150	-0.0003	-0.0010	0.0140	0.1542
Year Four					
-0.1188	-0.0132	0.0000	-0.0002	0.0132	0.2600

Note. Positive values would indicate relatively better performance on anchor items. Negative values would represent relatively better performance on nonanchor items.

Table 4.

Distribution of Difference in Average Person Residual (Anchor – Unique) by Grade

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Grade 3					
-0.1227	-0.0120	0.0004	0.0000	0.0123	0.1542
Grade 4					
-0.1444	-0.0154	0.0002	-0.0002	0.0156	0.1312
Grade 5					
-0.1509	-0.0153	-0.0003	-0.0002	0.0147	0.1344
Grade 6					
-0.1362	-0.0134	0.0004	0.0002	0.0139	0.1210
Grade 7					
-0.1166	-0.0132	0.0000	-0.0002	0.0125	0.2600
Grade 8					
-0.1321	-0.0136	0.0007	-0.0004	0.0142	0.1220
Grade 11					
-0.1565	-0.0129	-0.0003	-0.0013	0.0117	0.1388

Note. Positive values would indicate relatively better performance on anchor items. Negative values would represent relatively better performance on nonanchor items.

Table 5.

Distribution of Difference in Average Person Residual (Anchor – Unique) by Subject

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
		Mathematics			
-0.1557	-0.0126	0.0004	0.0000	0.0130	0.1368
		Reading			
-0.1509	-0.0148	0.0000	-0.0002	0.0144	0.2600
		Science			
-0.1565	-0.0149	0.0002	-0.0011	0.0144	0.1388

Note. Positive values would indicate relatively better performance on anchor items. Negative values would represent relatively better performance on nonanchor items.

Table 6.

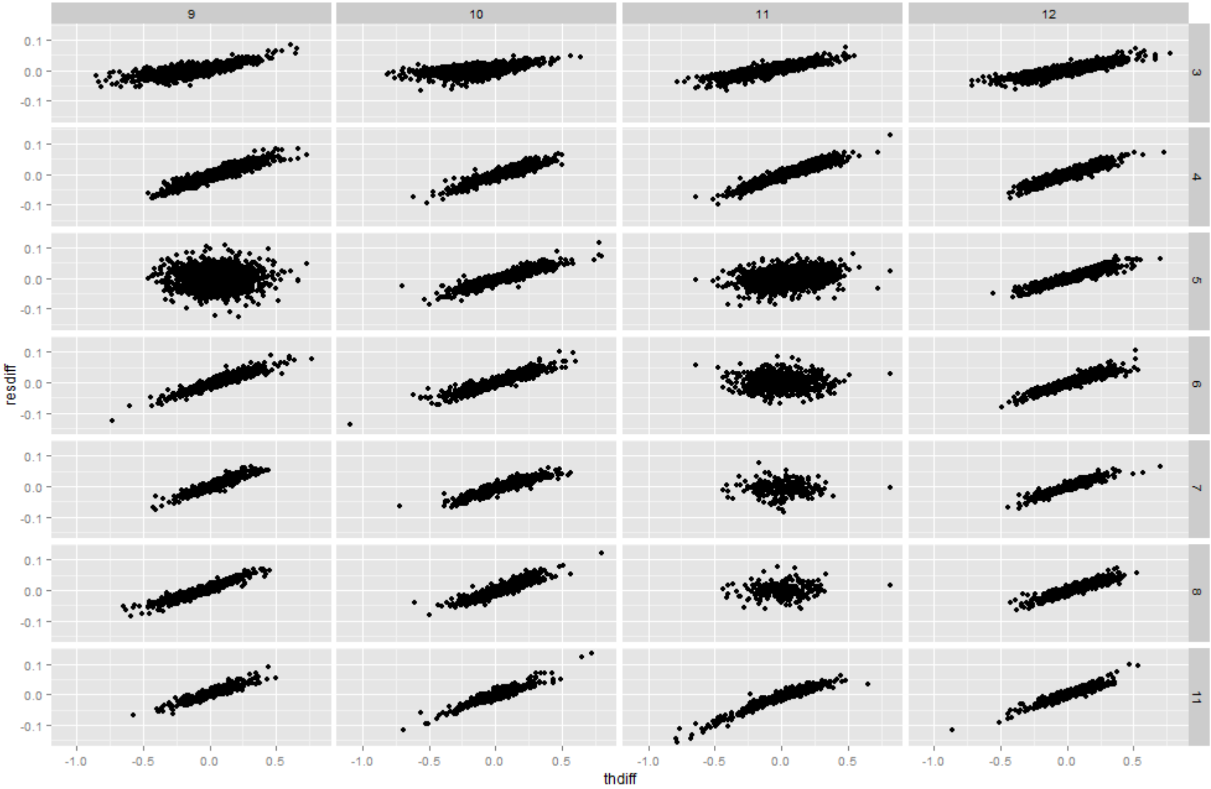
Number of Schools Flagged More than Three Times Across all Years, Grades, and Subjects

Flagging Criteria	Number of Schools
Theta difference > 3 SD units	11
Theta difference < -3 SD units	16
Residual difference > 3 SD units	15
Residual difference < -3 SD units	17

Note. Positive values would indicate relatively better performance on anchor items. Negative values would represent relatively better performance on nonanchor items.

Figure 1.

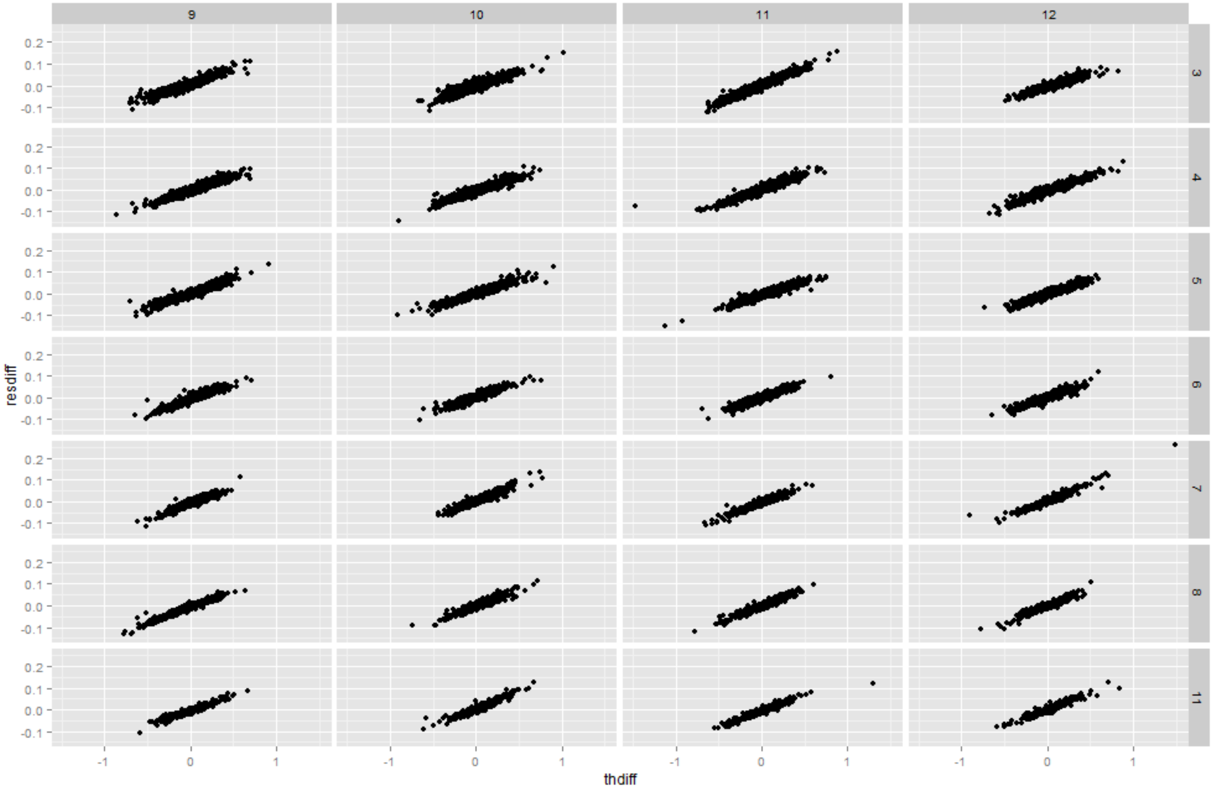
Residual difference (anchor – unique) plotted on theta difference (anchor – unique): Math



Note. Positive values would indicate relatively better performance on anchor items. Negative values would represent relatively better performance on nonanchor items.

Figure 2.

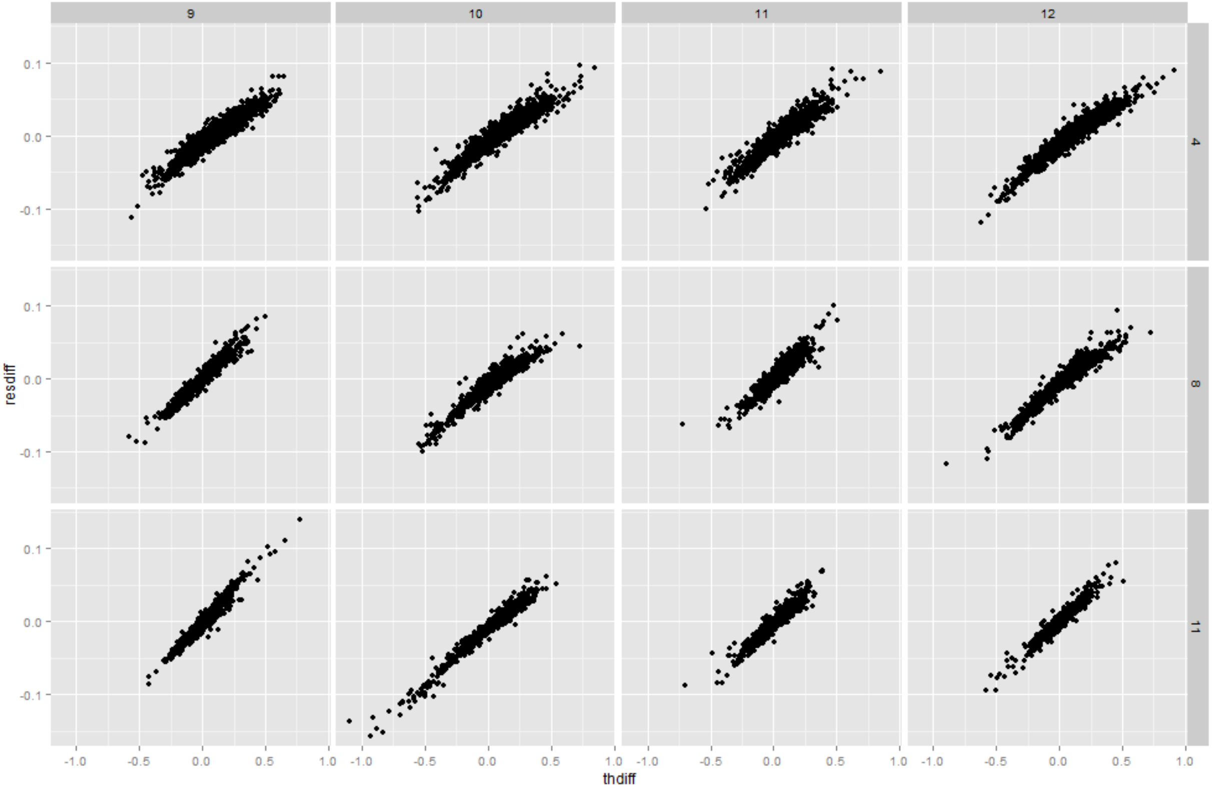
Residual difference (anchor – unique) plotted on theta difference (anchor – unique): Reading



Note. Positive values would indicate relatively better performance on anchor items. Negative values would represent relatively better performance on nonanchor items.

Figure 3.

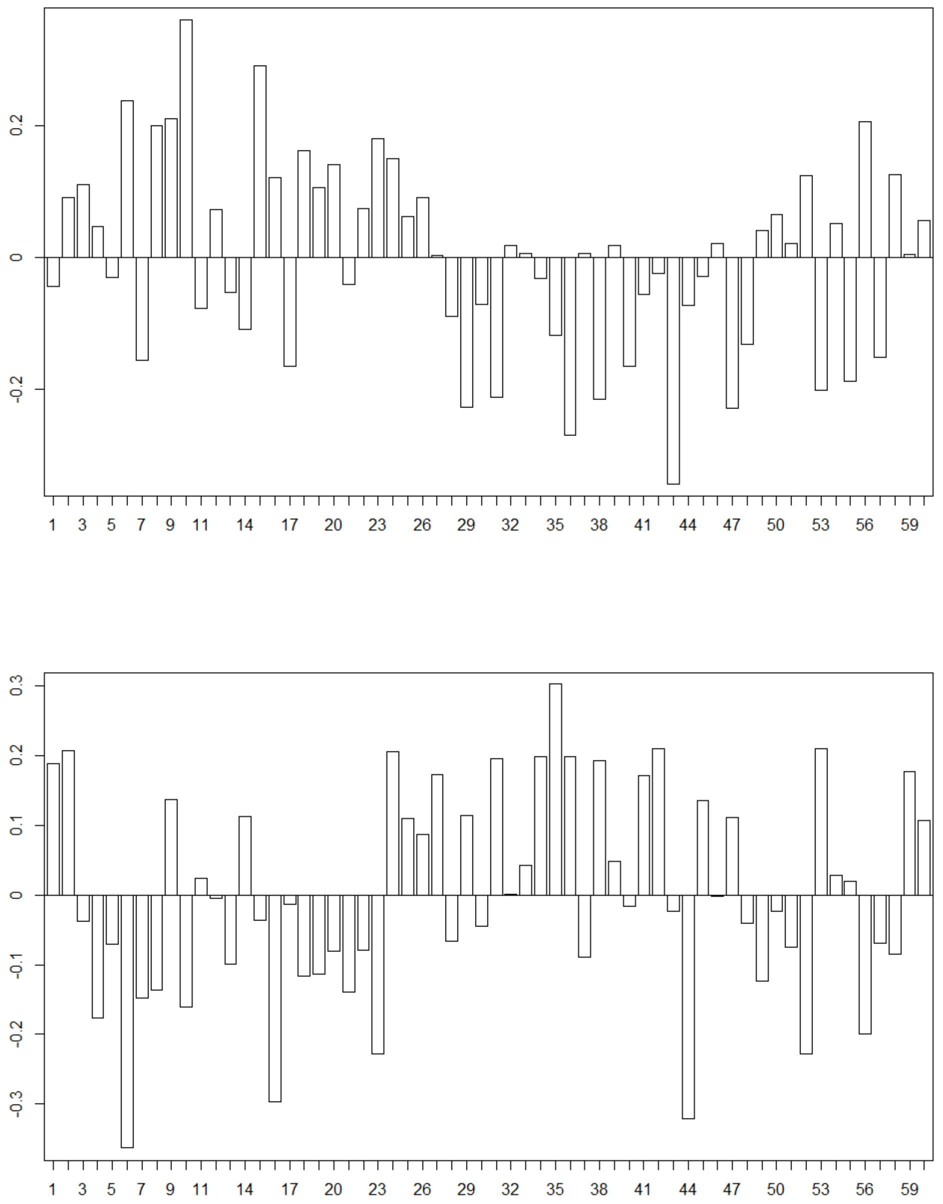
Residual difference (anchor – unique) plotted on theta difference (anchor – unique): Science



Note. Positive values would indicate relatively better performance on anchor items. Negative values would represent relatively better performance on nonanchor items.

Figure 4.

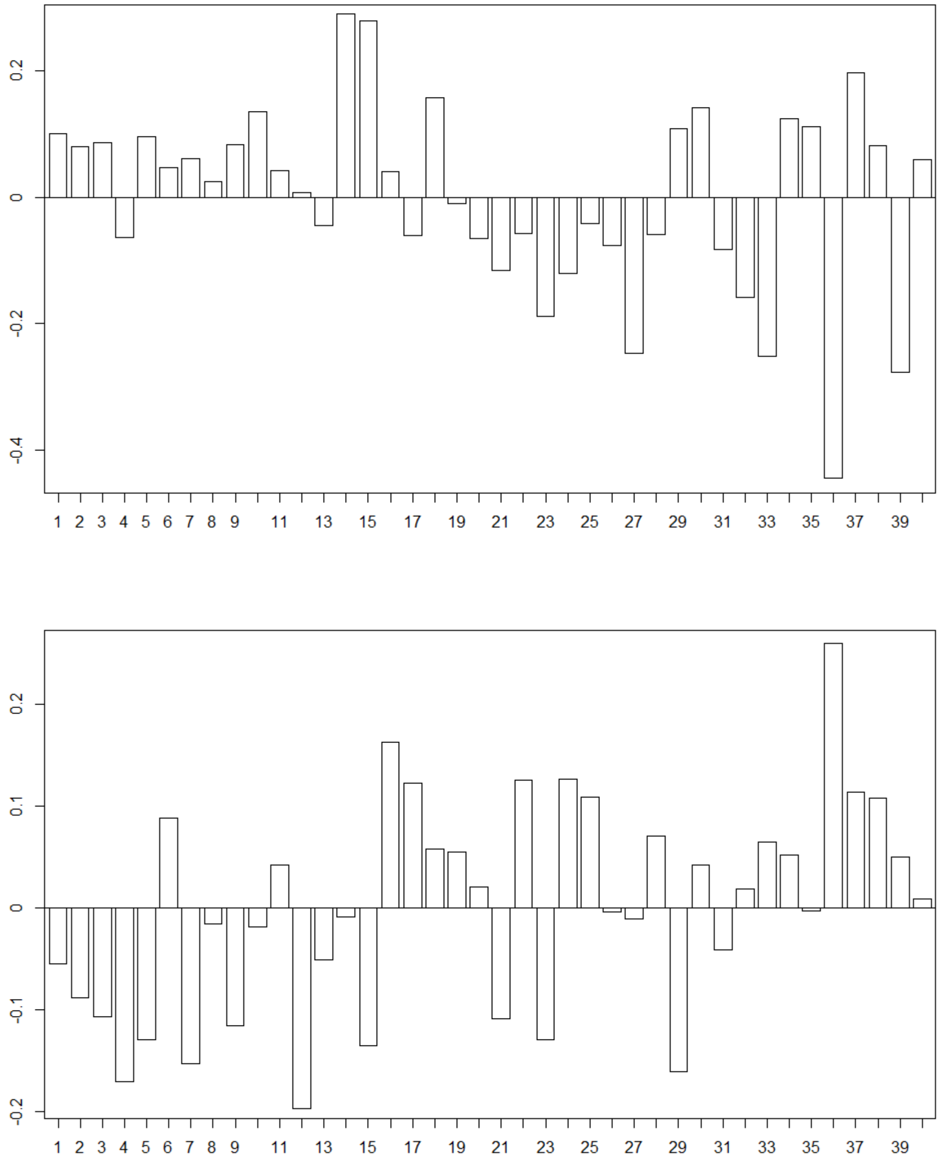
Grade 4 Math classes with the largest positive and negative residuals differences



Note. The first 24 items are anchor items. The remaining items are nonanchor items.

Figure 5.

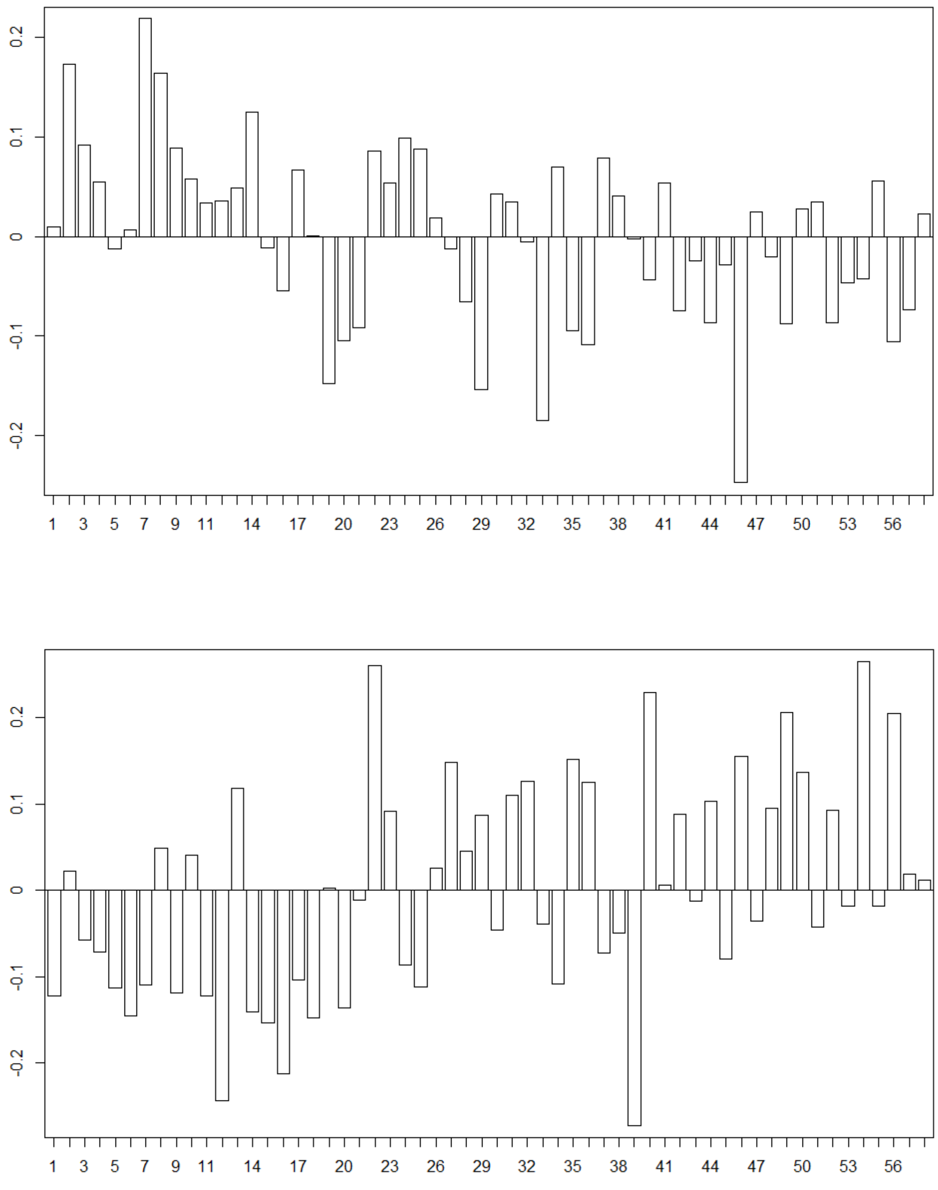
Grade 4 Reading classes with the largest positive and negative residuals differences



Note. The first 15 items are anchor items. The remaining items are nonanchor items.

Figure 6.

Grade 4 Science classes with the largest positive and negative residuals differences



Note. The first 16 items are anchor items. The remaining items are nonanchor items.