C-R-E-S-S-T

National Center for Research on Evaluation, Standards, and Student Testing

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Omitted and Not-Reached Items in Mathematics in the 1990 National Assessment of Educational Progress

CSE Technical Report 357

Daniel Koretz and Elizabeth Lewis
The RAND Corporation

Tom Skewes-Cox and Leigh Burstein
Center for Research on Evaluation, Standards, and Student Testing (CRESST)
University of California, Los Angeles

January 1993

Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported under the National Center for Education Statistics Contract No. RS90159001 as administered by the U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the National Center for Education Statistics or the U.S. Department of Education.

OMITTED AND NOT-REACHED ITEMS IN MATHEMATICS IN THE 1990 NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Daniel Koretz and Elizabeth Lewis
The RAND Corporation

Tom Skewes-Cox and Leigh Burstein
Center for Research on Evaluation, Standards, and Student Testing (CRESST)
University of California, Los Angeles

Unplanned non-response to cognitive items on the National Assessment of Educational Progress has been a concern for some time, particularly in the case of mathematics. Until recently, the primary concern has been "not-reached" items—that is, items not answered because the student failed to complete one or more of the blocks of items out of which the test booklets are constructed. The 1986 NAEP mathematics assessment was highly speeded and had serious problems with not-reached items. Some not-reached items were excluded from scaling, but the problem was so severe that items with not-reached rates up to .45 were scaled. Of the 446 unique items, 79 (18%) were not scaled because their not-reached rates exceeded .45. An additional 104 items (23%) were scaled even though they had weighted not-reached rates above 0.20 (Johnson, 1988).[1]

Although the speededness of mathematics blocks was greatly reduced in the 1990 mathematics assessment, ongoing changes in the format of the assessment have raised additional issues of non-response. The 1990 mathematics assessment includes increasing numbers of open-format items of different sorts, some of which require substantial constructed responses, such as explaining the solution to a problem. A number of observers have expressed concern that the rate of non-response to these items might be either unacceptably high overall or too disparate across groups of students. In addition, difficult open-ended items

[1] **These proportions, and all other results unless otherwise noted, are weighted using the overall student full sample weights.**

might cause some students to give up part way through a block, causing unacceptably high non-response rates to subsequent, traditional items as well.

Questions Addressed in This Study

To address these concerns, we examined patterns of non-response in all three age/grade groups (age 9/Grade 4, age 13/Grade 8, and age 17/Grade 12) included in the 1990 assessment of mathematics. Our analysis used the main sample that was the basis of the primary reported cross-sectional results of the assessment.[2] The tests used in the main sample are constructed of three short "blocks" of items; these blocks are then built into test booklets (with block position varied from booklet to booklet). We examined non-response rates for every item in the seven blocks of items upon which the reported scaled scores were based. In addition, we conducted limited analysis of three experimental blocks (comprising estimation items and items tapping higher-order thinking skills) that were not used to construct scaled scores.

As a first step, we examined non-response patterns for each item in each scaled block. The relationship of non-response to the position of blocks within booklets was investigated to determine whether more detailed analysis should be conducted on a pooled or a within-position basis.

As a second step, we examined differences in non-response for population groups (blacks, Hispanics, and non-Hispanic whites) and gender. Simple differences among population groups were explored by two further analyses. First, for a limited number of items, we matched white and black students in terms of percent correct on a mini-test comprised of items for which non-response rates were low, and then we compared group differences in non-response for these matched groups to the simple group differences. This provides some information on the extent to which group differences in non-response reflect disparities in mathematics proficiency. Second, we created subsamples that included (a) for schools with more tested whites, all tested black students but a random sample of whites equal in number to the tested blacks, and (b) for schools with more tested blacks, all tested white students but a random sample of blacks equal in number to the number of tested whites.

[2] The National Assessment includes a number of different samples in each age group. For example, trend estimates and cross-sectional estimates are based on different samples and partially different item sets.

Aggregating these subsamples provided a test of the extent to which white/black differences in non-response are related to variables associated with schools and communities.

Finally, we explored the relationship of non-response to characteristics of items. The characteristics we explored include item format, cognitive complexity, and the a priori content area classifications of the National Assessment of Educational Progress (NAEP).

Categories of Non-Response

As suggested above, there are two types of non-response in the NAEP: items can be omitted or not reached. Items are classified as omitted if a student resumes responding to later items within the block and are classified as not-reached if a student does not respond to later items. Unfortunately, although blocks are rotated through positions within booklets, items are not rotated across positions within blocks. When an item near the end of a block has a high not-reached rate and follows one or more other items with high not-reached rates, one cannot infer what response pattern the item would have generated had it been placed earlier in the block, because there are no blocks in the NAEP that include that item in a different position.

This feature of the NAEP design limits what can be inferred about the correlates of non-response. In the case of items that are preceded by others with high not-reached rates, it is always unclear whether the focal item's not-reached rate stems from its characteristics, its position, or both. Such items may be important nonetheless; for example, it would be important if a certain type of item happened to have a high non-response rate even if the reason were only the fact that such items tended to be at the end of blocks. Patterns of non-response among items bracketed by others with high rates of appropriate response, however, are more informative.

For this reason, we also created a "change in not-reached" (CNR) category comprising items that are accompanied by a substantial increase in the not-reached rate. In theory, such items could be informative, but we found relatively few instances in which the increase in the not-reached rate was uneven enough to distinguish individual items in this way.

Overall Severity of Omit and Not-Reached Rates

An item was considered to have a high omit rate if more than 10% of the students in a grade-only sample who received the item omitted it but responded to subsequent items. A 15% cutoff was used to define items as having high not-reached rates. For reasons that are described in a subsequent section, the overall rates presented here are pooled across all block positions.[3]

Omit Rates

The problem of omitted items was appreciable in Grade 12, in which 9% (13 of 144) of the items had omit rates exceeding our threshold of 0.10. The problem was less severe, however, in the younger grades. Only 5% (5 of 109) of the items in Grade 4 and 6% (8 of 137) in Grade 8 had omit rates exceeding .10.

Not-Reached Rates

The problem of large not-reached rates was markedly less severe in the 1990 assessment than in 1986. As noted earlier, in the 1986 mathematics assessment, 41% of the items administered (23% of the items finally scaled) had not-reached rates above 0.20. In contrast, only 23 (8%) of the 275 math items which were scaled in the 1990 assessment had not-reached rates above 0.20 (Table 1).[4] Nonetheless, some blocks in the 1990 assessment still had a substantial problem with not-reached items, and the highest not-reached rate was 0.45.

Items Causing Large Changes in Not-Reached Rates

As noted above, the structure of NAEP obscures the reasons that some items have high not-reached rates. Because items are not repeated in multiple positions within blocks, one cannot tell whether an item has a high not-reached rate because of position or because of item characteristics. In an attempt to disentangle these factors, we looked for items responsible for a sizable increase

---

[3] We tried a number of other cut-points as well and found that the general conclusions reported here are fairly robust across moderate differences in cut-points. For example, the general pattern of population-group differences in omit and not-reached rates was reasonably stable across the majority of cut-points we tried.

[4] The main (BIB) sample included 282 items. Seven were not scaled, but they were not dropped because of high not-reached rates. They were dropped either because their estimated item response functions were non-monotonic or because the data analysis, statistics, and probability subscale was dropped for Grade 4 (see Yamamoto & Jenkins, 1992).

Table 1

Scaled Items With High Not-Reached Rates, by Grade

	Grade 4	Grade 8	Grade 12
Not-reached rates ≥ .15	9	7	13
Not-reached rates ≥ .20	6	6	11
Not-reached rates ≥ .30	2	2	5
Not-reached rates ≥ .40	1	0	2

in the not-reached rates, flagging all items that were accompanied by a change in the not-reached rate (CNR) of at least 0.05.

Few items had high CNR rates, however. In general, once not-reached rates began rising as students progressed through a block, they rose gradually, and the change in not-reached rates from one item to the next was generally small. Only 4 of 109 scaled items in Grade 4 and only 4 of 137 items in Grade 8 exceeded a CNR threshold of 0.05. In Grade 12, 13 of 144 scaled items had CNR rates over 0.05.

Characteristics of Problematic Items

Non-response rates were related to characteristics of items, although not always in the ways anticipated.

Block Position

The scaled NAEP mathematics booklets consisted of background questions followed by three blocks of mathematics questions. These three blocks were separately timed, with an equal amount of time devoted to each block. A total of seven blocks of questions were scaled, and each of the seven occurred in each of three possible within-booklet positions. We expected that non-response would sometimes worsen from position 1 to position 3 because of growing fatigue.

The data, however, generally did not conform to our expectations; in all three grades, we found that block position generally made little or no difference in the non-response rates. In a few instances, omit rates were higher in position 3, but the differences in rates across positions were very small. Moreover, in the

case of not-reached rates, the few differences that emerged countered our expectations: not-reached rates were marginally higher in position 1.[5] In all cases, the effects of position were so small that subsequent analyses were carried out by pooling results across positions.

Given that position effects were infrequent, small, and in opposite directions for omit and not-reached rates, it would be risky to place much confidence in any interpretation of them. One could argue that two factors are likely to be at work. Fatigue might increase non-response from position 1 to position 3. On the other hand, practice effects might decrease non-response, as students learn in responding to the first or second block that they need to pace themselves differently to complete the block. It is not apparent, however, why these two factors would have differential effects on omit and not-reached rates.

Item Format

Virtually all of the items that had omit rates above 0.10 had some type of open-ended format. This was true of all 5 Grade-4 items, all 8 Grade-8 items, and all but 1 of the 13 Grade-12 items with omit rates above 0.10.

The high omit rates, however, appear to be a function of difficulty as well as format. Only a small minority of the open-format items had high omit rates, particularly in Grades 4 and 8 (see Table 2). The open-ended items that have high omit rates are, on average, an atypically difficult subsample of the open-ended items, and they are more difficult than the average multiple-choice items as well.[6]

Not-reached rates show a far less clear relationship to item format. Out of 29 items with not-reached rates over 0.15 in any one of the three grade-only samples, only 11 were open-ended. This weak relationship, however, may be an

---

[5] Some of the position differences are larger for blacks.

[6] The difficulty of items with high omit rates is unclear because there is no certainty about the likelihood that students who omitted the items would have answered them correctly if they had attempted to. The *p*-values in Table 1 were computed by ETS as the number of correct responses divided by the sum of the correct, incorrect, omitted and multiple responses. (Students who did not reach a particular item were excluded from the calculation. See Beaton & Zwick, 1990, p. 66.) This method counts omitted items as wrong and is equivalent to assuming that students who omitted the item would have been unable to answer it. We also computed *p*-values as the number of correct responses divided by the sum of correct, incorrect and multiple responses and found that although this slightly narrows the difference between the *p*-values, the conclusions are not altered appreciably.

Table 2

Mean *p*-Values for Multiple-Choice Items, All Open-Ended Items, and Items With High Omit Rates

	All multiple-choice		All open-ended		High omit rates	
	p-Value	N	*p*-Value	N	*p*-Value	N
Grade 12	.58	109	.44	35	.23	13
Grade 8	.57	102	.50	35	.28	8
Grade 4	.53	81	.49	28	.23	5

artifact of the structure of blocks. Not-reached items are by definition at the end of each block, and the majority of items in that position are multiple-choice.

Items with high not-reached rates, like those with high omit rates, tend to be more difficult than others. For all three of the grade-only samples, the average *p*-value for the items with high not-reached rates is .27, compared to .56 for all other items.[7] This relationship, however, is also clouded by the structure of blocks, because items at the ends tend by design to be difficult.

We did not find any clear-cut relationship between item format and CNR rates, in part because it is not clear whether items have high not-reached rates because of within-block position or other characteristics. Thus, even though all four Grade-4 items associated with large increases in not-reached rates were open-ended items, the impact of format cannot be ascertained. In Grade 12, 6 of the 13 items with high CNR rates were open-ended.

Item Classifications

In addition to the block position and item format we considered two additional item characteristics: content area and cognitive operations.

Content area. ETS classified items into five major content areas: (a) numbers and operations; (b) measurement; (c) data analysis, statistics and probability; (d) geometry; and (e) algebra and functions. Each of these areas was divided into topics and each topic was further divided into subtopics. The total number of subtopics was 153.

[7] There was almost no variation in these numbers across grades.

We were concerned that ETS's basic content classification might have been too coarse to identify differences in non-response and that the more detailed topic and subtopic classifications yielded too sparse a matrix. (The 153 subtopics compare to 275 unique scaled items in the main assessment for Grades 4, 8 and 12.) Accordingly, we developed our own content classification scheme with eight content areas: (a) computation/miscellaneous; (b) fractions/decimals/percents; (c) algebra; (d) mensuration; (e) tables/graphs; (f) geometry/trigonometry; (g) measurement; and (h) probability/statistics. We applied our alternative classification to the Grade-8 sample.

Average overall omit rates (that is, across all students) varied neither markedly nor consistently across the ETS content areas in any of the three grades (Table 3). The mean omit rate for numbers and operations items was relatively low in all grades, and the rate for measurement items was relatively low in Grades 4 and 8. The inconsistency of the differences across grades, however, suggests that apart from numbers and operations, differential omit rates are more a function of the specific items used in each grade rather than of content area *per se.*

Our more detailed analysis of content classifications in Grade 8 also showed relatively little in the way of differences across the eight content areas. In proportional terms, the data analysis, statistics and probability area had a larger proportion of items with high omit rates than the other areas using both our classification and ETS's, but this reflects so few items (2 out of 12 and 2 out of 19, respectively, in that area) that it could well be chance and may not reflect anything specific to that content area.

Table 3

Mean Omit Rates by Grade and ETS Content Area

	Numbers and operations		Measurement		Geometry		Data analysis		Algebra	
	Omit rates	No. of items	Omit rates	No. of items	Omit rates	No. of items	Omit rates	No. of items	Omit rates	No. of items
Grade 4	2.5	52	2.0	20	5.2	14	3.9	9	4.6	14
Grade 8	2.7	46	1.4	21	2.5	26	3.7	19	2.9	25
Grade 12	2.0	37	4.6	23	3.0	25	3.4	22	4.2	37

Mean not-reached rates varied more sharply among the ETS content areas, but still inconsistently across grades (Table 4). Atypically high average rates appear in geometry in Grade 4 and in algebra in Grade 12. Because not-reached rates are determined by the structure of blocks, these patterns need not indicate any differences among content areas *per se,* but even if they are idiosyncratic results of the construction of this particular assessment, they illustrate non-response problems that should be routinely monitored.

Cognitive operations. ETS also classified items into three learning technique areas: Conceptual Understanding, Procedural Knowledge and Problem Solving. These three areas were further subdivided into a total of 16 subareas. We were concerned that an additional dimension, cognitive complexity, might contribute to non-response problems, so we developed an alternative scheme of cognitive complexity which broke down the items into four classes: (a) problems involving a single step, procedure, or concept; (b) real-world, multiple-step problems; (c) standard classroom, multistep problems; and (d) novel problems (see Appendix).

Neither cognitive-operations classification was strongly related to omit or not-reached rates. Our complexity dimension was decidedly skewed, however, with over 80% of the items falling in the first category (one-step problems involving the application or recall of a single procedure or principle). The low counts in the remaining categories preclude any firm interpretation of contrasts among the groups.

Table 4

Mean Not-Reached Rates by Grade and ETS Content Area

	Numbers and operations		Measurement		Geometry		Data analysis		Algebra	
	Not-reached rate	No. of items	Not-reached rate	No. of items	Not-reached rate	No. of items	Not-reached rate	No. of items	Not-reached rate	No. of items
Grade 4	3.5	52	4.9	20	7.3	14	3.6	9	3.8	14
Grade 8	4.2	46	3.5		2.0	26	2.8	19	3.0	25
Grade 12	1.7	37	2.6	23	2.8	25	3.4	22	9.9	37

Differences Among Groups of Students

A primary motivation for this study was concern about possible differential non-response across population groups, but we also examined gender differences.

Gender Differences in Omit and Not-Reached Rates

No items showed large gender differences in not-reached rates, but a few showed sizable differences in omit rates. Five items in Grade 12 and two in Grade 4 showed sizable gender differences in omit rates. There were no major gender differences for Grade 8; all of the observed gender differences in that grade were less than three percentage points. The significance of these differences, already suspect because of the small number of items involved, is thrown further into doubt by inconsistent sign: The female omit rates were higher for the five Grade 12 items, but the male omit rates were higher for the two Grade 4 items. With the exception of two items in the Grade 12 sample, the items showing gender differences in omit rates were open-ended.

The few gender differences in omit rates in Grade 12 appear to be at least partially independent of proficiency differences between males and females. To explore this question, we performed a Mantel-Haenszel (MH) analysis in which the outcome was omit rates rather than the more common *p*-values.[8] Six proficiency intervals were used, and the cut-scores were selected to yield roughly equal frequencies in all six.[9] All tabulations were run without weights and with not-reached treated as missing; tabulations of gender differences (in Block 9, one of several that included items with substantial nonresponse rates) showed that these decisions had essentially no impact on the relative percentages in the 2x2x6 contingency table created for the MH procedure.

The MH analysis suggests that Grade-12 gender differences in omit rates cannot be fully explained by proficiency differences. Of the five items that had omit-rate differences of at least 0.05 in Grade 12, three had sizable and highly significant MH chi-squares, ranging from roughly 16 (p=.009 after a Bonferroni correction for 144 multiple comparisons) to 51.[10] A fourth item had a chi-square of 8, which corresponds to an unadjusted probability of .005 but becomes

---

[8] This analysis was suggested by Ed Haertel.

[9] The final counts in the six intervals for 12th-grade students responding to Block 9 ranged from 354 to 530; four were between 416 and 487.

[10] These analyses were conducted without correction for continuity.

nonsignificant after adjustment for multiple comparisons. (Note that a Bonferroni correction for 144 multiple comparisons is a very conservative approach to item-wise significance levels.) The fifth item produced a small MH statistic (chi square = 3.5) that was not significant even before adjustment.

The MH analysis also suggests that a simple comparison of arithmetic differences in omit rates may not always fully capture differences among groups. Our analysis revealed an additional 26 items which had failed to meet our criterion of a 0.05 difference in omit rates but nonetheless yielded MH chi-squares with unadjusted probabilities below .05. Many of these chi-squares were large, but only two remained significant after adjustment for 144 multiple comparisons. One factor that could contribute to this finding is that the distribution of percentages is compressed at the tails, so when both groups have relatively low omit rates, the arithmetic difference between them is necessarily small.[11] In many cases, the within-group omit rates are low enough that the differences among groups may be of no practical importance, but further exploration is needed to determine what other conditions may obscure differential omitting when simple mean differences in rates are not large.

Population-Group Differences in Omit and Not-Reached Rates

We compared omit and not-reached rates for the three largest population groups represented in the assessment (white, black, and Hispanic).

Hispanics and blacks have higher average omit rates than whites in all grades (Table 5). Underlying this small but consistent difference in means are substantial group differences in omit rates on a smaller number of items. The number of items involved of course depends on the size of the difference in omit rates used as a criterion: The larger the difference used as a cut-score, the fewer items show differential omit rates (Table 6). Six to 15% of all items showed a 5-percentage-point difference in omit rates between whites and either blacks or Hispanics. The proportion of items reaching this level of difference between groups was lower in Grade 4 than in the other grades.

Most of the items showing black-white or Hispanic-white differences in omit rates of at least 0.05 were open-ended. The exceptions are three Grade 12

---

[11] Examination of the four items that had the largest MH chi-squares but failed to reach a difference of 0.05, for example, showed one just missed that criterion, and the other three all had low within-group omit rates, ranging from 0.007 to 0.07.

Table 5

Mean Omit and Not-Reached Rates by Population Group and Grade

	Omit Rates			Not-Reached Rates		
	Whites	Blacks	Hispanics	Whites	Blacks	Hispanics
Grade 4	2.8	4.3	4.0	3.6	6.9	5.5
Grade 8	2.1	4.0	3.9	2.6	5.6	4.8
Grade 12	2.9	4.8	4.9	3.7	7.2	6.7

Table 6

Population-Group Differences in Omit Rates (Percentage of the items at each grade level which exceed the stated group differences in omit rates)

Size of group difference	Grade 4		Grade 8		Grade 12	
	White-Black	White-Hispanic	White-Black	White-Hispanic	White-Black	White-Hispanic
0.050	8	6	12	15	11	12
0.075	5	5	5	5	6	9
0.100	4	2	2	3	3	6
0.150	3	1	0	0	1	2
0.200	1	0	0	0	0	0
Number of items	109		137		144	

multiple-choice items with large Hispanic-white differences (of 5 to 8 percentage points).

The population-group differences in omit rate did not show striking or consistent relationships with content area. Geometry showed somewhat larger than average mean differences among population groups in Grade 4 but not in Grades 8 or 12 (Table 7). Mean differences were relatively large in data analysis, probability, and statistics in all grades, but the disparity between this content area and the others was modest.

In Grade 8, we approached this question also by tabulating the number of items that had omit rates above 0.10 within population group and content

Table 7

Mean Population-Group Differences in Omit Rates by Grade and ETS Content Area

	Numbers and operations		Measurement		Geometry		Data analysis		Algebra	
	Black-White	White-Hispanic	Black-White	White-Hispanic	Black-White	White-Hispanic	Black-White	White Hispanic	Black-White	White Hispanic
Grade 4	1.0	0.7	0.9	1.0	3.9	3.0	2.7	2.6	1.2	1.0
Grade 8	1.5	1.3	1.1	1.3	1.8	1.6	3.1	3.0	2.6	2.4
Grade 12	1.6	1.4	1.6	1.7	1.8	1.8	3.2	3.6	1.7	1.8

classification. By this standard, the population group difference in data analysis was not atypical. Data analysis had the highest proportion of items with high omit rates in all content areas (Table 8), but the population-group differences were proportionally as large in numbers and operations and algebra. (That is, the ratios of the number of items with omit rates above 0.10 for blacks to the number with such rates for whites were similar.)

Mean population-group differences in not-reached rates were somewhat larger than differences in omit rates (Table 5), and a larger proportion of items showed sizable differences in not-reached rates (Table 9). However, the items showing differential not-reached rates, like those showing high overall not-

Table 8

Number and Percent of Items With Omit Rates Above 0.10, by Population Group and ETS Content Area, Grade 8

Areas	Whites		Blacks		Hispanics	
	Number	Percent	Number	Percent	Number	Percent
Numbers and operations	2	4	5	11	5	11
Measurement	0	0	0	0	1	5
Geometry	1	4	1	4	2	8
Data analysis	2	11	6	32	6	32
Algebra	0	0	4	16	4	16

Table 9

Population-Group Differences in Not-Reached Rates (Percentage of the items at each grade level which exceed the stated group differences in not-reached rates)

Size of group difference	Grade 4		Grade 8		Grade 12	
	White-Black	White-Hispanic	White-Black	White-Hispanic	White-Black	White-Hispanic
0.050	20	12	22	17	27	24
0.075	14	5	15	11	19	17
0.100	8	2	12	7	13	12
0.150	3	0	3	3	2	3
0.200	0	0	0	0	0	0

reached rates, had no particular characteristics other than difficulty and positions near the ends of blocks.

To determine the influence of proficiency differences on black-white differences in omit rates in Grade 12, we repeated the Mantel-Haenszel approach described for gender differences above. We again computed two probability values for each MH chi-square: unadjusted, and adjusted for 144 multiple comparisons.

Regardless of adjustment for multiple comparisons, the items selected by the MH procedure showed very limited overlap with the set selected because of having a black-white omit-rate difference greater than 0.05. A total of 16 out of 144 items scaled for Grade 12 had black-white differences in omit rates greater than 0.05 (Table 10). Far more items—38, about a fourth of the item set—had an unadjusted $p<.05$ from the MH analysis. Only 7 items, however, met both criteria. Only a small number of items still had statistically significant chi-squares after adjustment for multiple comparisons, and only about half of those also had arithmetic differences in omit rates greater than 0.05.

These results suggest that a substantial share of the black-white differences in omit rates can be accounted for by differences in proficiency. (Recall that in the case of gender differences, most of the items selected on the basis of differences in rates also had a significant MH chi-squares before adjustment.) As in the case of gender differences, many of the items that yielded large MH

Table 10

Grade-12 Items With Black-White Differences in Omit Rates: Arithmetic Difference and Unadjusted and Adjusted Mantel-Haenszel Chi-Squares

Alpha for MH Chi-Squares	Omit-rate difference > .05	Significant MH chi-square	Both criteria
No adjustment, $p<.05$	16	38	7
Bonferroni adjustment, $p<.10$	16	9	4
Bonferroni adjustment, $p<.05$	16	5	3

statistics but small arithmetic differences had low within-group omit rates. Here again, however, further exploration of group differences that may be obscured by simple mean differences seems warranted.

As a second, exploratory approach to the question of the impact of proficiency differences on black-white differences in omit rates, we compared matched subsamples of black and white 12th-grade students. We matched white and black students who had received math block M9 in terms of their percent-correct scores on a small subtest comprising 12 of the 20 items from that block. Items were included in the subtest only if the sum of the omit and not-reached rates was less than or equal to 0.15. The population-group differences in non-response rates for these matched groups were compared to the simple group differences in the unmatched sample. The analysis focused on the five items in block M9 with the largest white-black differences in omit and not-reached rates in the unmatched sample.

This approach was followed for only a single block, and the findings for the few items in that block with high omit rates may not be indicative of patterns in the assessment as a whole. It is noteworthy, however, that the results are basically consistent with the MH analyses already reported. For four of the five items examined, the white-black difference in omit rates for the matched sample was considerably smaller than the difference in the total sample (Table 11). This implies that a substantial portion of the higher omit rate for black students can be attributed to proficiency differences. In addition, we plotted the omit rates separately for whites and blacks in the matched samples at each percent-correct

Table 11

Black-White Differences In Omit Rates For The Total And Ability-Matched Samples. (Listed in order of overall white-black difference.)

	Total sample			Matched sample		
Item number	White	Black	Black-White	White	Black	Black-White
9	20.8	28.6	7.9	24.4	27.9	3.5
8	7.3	11.7	4.4	8.3	11.1	2.8
12	5.5	9.3	3.7	7.9	8.1	0.3
18	12.5	15.0	2.5	15.5	13.7	-1.8
3	4.5	6.8	2.3	5.1	7.5	2.5

score on our subtest. The small numbers involved (particularly for blacks at the higher scores) and the instability of the rates across groups precludes reaching any firm conclusions, but it appears that although the omit rate is higher for students of either race with low scores on our subtest, the size of the white-black difference in omit rates does not change consistently as scores rise. Because proportionally more blacks are in the lower scoring groups, this pattern is also consistent with the conclusion that only part of the omit-rate difference on those items can be attributed to proficiency differences.

The white-black difference in not-reached rates was smaller for the matched sample for the first two items and approximately equal for the third item (Table 12).[12] For the last two items, the white-black differences were *larger* in the matched sample. However, these last items were at the end of the block, and, as noted earlier, performance on those items is difficult to interpret.

The results of both the matching and MH analyses are reasonably consistent with the results of Swinton's (1991, 1992) analysis of black-white differences in response rates, which used both different samples and different methods. Swinton's work applied regression models to limited samples of items to partial out the effects of proficiency on non-response. Swinton found that "ethnic difference in responding can be accounted for to a great extent by ability

[12] Note that four of the five items in Table 12 are algebra items, while only one of the items in the subtest used for matching was an algebra item.

Table 12

Black-White Differences in Not-Reached Rates for the Total and Ability-Matched Samples (Listed in order of overall white-black difference)

Item number	Total sample			Matched sample		
	White	Black	Black-White	White	Black	Black-White
18	21.8	37.1	15.4	22.2	38.0	15.8
17	14.5	27.6	13.1	14.9	26.9	12.0
19	31.3	44.7	13.3	29.7	45.5	15.9
16	11.2	24.0	12.8	11.7	22.8	11.1
20	38.4	50.1	11.6	35.7	52.1	16.4

difference" (Swinton, 1992). Partialing proficiency reduced the race main effect on omit rates among 17-year-olds (but not 13- and 9-year-olds) to statistical insignificance. However, item type still showed a large main effect, as well as a substantial interaction with race. Although Swinton's samples and methods differed from ours, his findings appear consistent with our finding that measured ability does not entirely remove the tendency for black students to have higher omit rates than whites on a limited number of items, most of which are open-ended.

In addition to the analysis described above, we explored the influence of variables associated with schools and communities on population-group differences in non-response. These variables, like directly measured proficiency (with which they are presumably highly correlated), appear to account for some but not all of non-response differences.

We began by creating a subsample of the Grade 12/age 17 sample with the same number of white and black students per school. If the school contained more tested white students, a random sample of whites equal to the number of tested blacks was selected; if the school contained more tested black students, a random sample of blacks equal to the number of tested whites was chosen. The

school-matched sample comprised 310 white and 310 black students in 114 schools.[13]

Black-white differences in omit rates were on average smaller in the school-matched sample (mean = 0.9) than in the total sample (mean = 1.7). The reason is that the omit rates for whites from the school-matched sample were on average larger by 1% than those from the total sample of whites. (The mean omit rate for blacks was similar in the school-matched and total samples.) A similar result is found with the not-reached rates: The average black-white difference for the total sample is 5.6, compared with 3.6 for the school-matched sample.

Conclusions and Recommendations

In the 1990 NAEP mathematics assessment, overall omit rates were modest in Grades 4 and 8, and not-reached rates were greatly reduced from 1986 levels. Differences in non-response between white and minority students were less severe than they first appeared, in that they appear to be partly attributable to proficiency differences. Gender differences in omit rates were infrequent.

Nonetheless, the results presented above provide grounds for concern. Omit rates were high for a subset of open-ended items, and the proportion of items with high omit rates in Grade 12 was substantial. The omit-rate differentials between white and minority students are troubling and will likely become more so as the NAEP continues to increase its reliance on open-ended items that black and Hispanic students currently show a greater propensity to omit. Not-reached rates remain high in the case of certain blocks.

Taken together, these results suggest the need for routine but focused monitoring and reporting of non-response patterns. Non-response problems may vary from one assessment to another, and there is some risk in generalizing too far from the present analysis of a single assessment. Absent comparable analyses of other assessments, however, the following guidelines seem warranted for monitoring non-response.

---

[13] The grade/age sample was used for this analysis because the subsamples created from the grade-only sample would have been too small.

Monitoring Overall Non-Response Rates

Omit rates and not-reached rates should be routinely monitored for all subject areas in each assessment cycle. The results of each assessment should also be screened for what we have called "high-CNR" items—that is, items that cause a sharp increase in not-reached rates. To the extent feasible, pilot studies should be used to reveal particularly severe problems of non-response.

Monitoring Differential Non-Response Across Item Types

The results of this study suggest that non-response rates should be monitored for specific categories of items as well as across the entire assessment. Monitoring focused on specific item types is needed to make certain that the potentially most severe problems of non-response are uncovered and addressed. This focused monitoring can also have a more positive function, however. As successive cohorts of students gradually gain more experience with problematic categories of items—say, those with particular formats or involving certain types of content or skills—rates of non-response to such items (or troublesome differences in rates among groups) may lessen. Routine, focused monitoring of non-response would reveal such improvements and may facilitate making desired changes in the assessment without generating undue response problems.

The results of this study suggest the following guidelines for focused monitoring of non-response:

Block position. To the extent that the 1990 mathematics assessment is typical, block position does not appear to be strongly related to non-response or to warrant extensive focused monitoring. However, given the possibility that mathematics may be atypical in this regard, cursory investigation of block position in other assessments would be prudent. Moreover, if the length of the testing sessions is increased, block position might well become more important.

Item format. Clearly, the relationship between item format and non-response must be monitored especially carefully, and rates for open-ended items must be checked in detail. The patterns we found in the 1990 mathematics assessment suggest that open-ended items are disproportionately likely to have high omit rates. Moreover, difficult open-ended items are more likely than others to cause high rates of non-response, but it remains unclear whether it was difficulty *per se* or some other attributes of those items (such as idiosyncrasies of

content or format) that accounted for their unusually severe non-response rates in the results reported above.

Item content. Although differential non-response across content areas was neither consistent nor particularly striking in the 1990 mathematics assessment, some exploration of differences across content areas probably should be included in routine monitoring of non-response. The hints of greater non-response in the data analysis, probability and statistics area suggest that items reflecting content that is inconsistently covered in schools may be a particular concern. The inconsistency of the differentials we found suggests that *a priori* content classifications may be insufficient basis for these investigations, however. In some instances, it may be appropriate to work backwards—for example, to examine items with large non-response differentials to see if some have common elements in terms of content or exposure that are not apparent from their *a priori* classifications.

Monitoring Differential Non-Response Across Groups

Our results indicate the importance of routine monitoring of non-response differentials across groups of students. Patterns in the 1990 NAEP mathematics assessment suggest that it is particularly important to examine differences among population groups. Although we found relatively few gender differences in non-response in the 1990 mathematics assessment, some degree of monitoring of gender differences would be sensible because gender differences in other subject areas might be larger. Indeed, a prudent course might be to conduct at least rudimentary checks of non-response differentials across variables that define NAEP's primary reporting categories.

Patterns in the 1990 mathematics assessment suggest that two further steps may be warranted where non-response differentials are found. First, analyses should be undertaken to explore the degree to which differential response may be attributable to overall proficiency differences. Second, exploration of interactions between item type and student characteristics—in particular, population group—appears warranted.

Explaining Failure to Reach Items: Varying Position

In a number of instances, our analysis could not clarify the correlates of high not-reached rates, which include some of the most striking instances of non-

response. This was because the within-block position of items was not varied. As a result, except in the case of the few "high-CNR" items, one cannot tell whether the high non-response rate is a function of position or other item attributes. This characteristic of the design is especially troubling now that the NAEP is moving toward increased use of open-ended items of the sort that appear to cause a greater propensity to omit among minority students.

There are at least two solutions to this problem, but both involve changing the design of the assessment. One approach would be to use extensive pilot testing to make sure that not-reached rates remain low throughout blocks. The advantages of this approach are apparent, but it has a number of potential drawbacks as well. It might require a time-consuming, iterative pilot testing, and it could produce unwelcome constraints on the construction of blocks. An alternative approach is to vary the position of a subset of items to ensure that response problems, if they arise, can be clearly identified. Items that are open-ended, particularly difficult, or involve novel content would all be candidates for varied positions if they are to be used toward the end of blocks. Even if this approach is followed, pilot testing could be used to narrow down the subset of items for which varying position would be worthwhile.

Reporting Non-Response

The severe non-response problems in the 1986 assessment, as well as the less severe but still troubling results reported here, argue that information on non-response should be reported routinely. However, the range of analyses suggested above would produce more information than would warrant reporting in documents for public consumption. This raises important questions of reporting: Which results should be reported routinely, which results warrant detailed presentation, and what vehicles should be used for reporting them?

A reasonably thorough overview of important non-response rates should be available to technical audiences. Moreover, it is important that the reporting of those rates be fairly consistent across subject areas and assessment cycles. A logical vehicle for that reporting would be the Technical Reports from the National Assessment of Educational Progress that are released in each National Assessment cycle. For example, a summary of the most important non-response findings could be included in the chapters devoted to the assessments in each subject area, and additional tabular detail could be provided as appendices. The

results reported through this vehicle should probably include rates by item type and population group, as well as overall rates. To the extent that potentially important information about non-response becomes too extensive for that vehicle, it could be included in the "almanacs." This would provide ready access for the small number of technically-oriented consumers who would want the information, particularly now that almanacs are provided on diskettes.

In addition, key findings about non-response may warrant inclusion in documents intended for wider consumption. These findings could be summarized in a format similar to (but presumably briefer than) that used for the procedural and data appendices included in recent *Report Cards*. Moreover, in certain cases, it may be important to include references to non-response in the main narrative of reports as well, in order to avoid misinterpretation. This might be important, for example, in cases where non-response is very high or where results are reported in forms (e.g., percentages of students providing adequate responses to open-ended questions, or *p*-values for multiple-choice items) that are sensitive to the treatment of non-response.

Reporting of non-response rates should not be limited to means. Although means or other similar statistics may be helpful for finding broad patterns in the data (such as the lack of a strong main effect of *a priori* content area on non-response rates), it is also important to report information at the item level. We recommend that reporting include the numbers and percentages of items exceeding specified omit and not-reached rates, as well as information describing those with high rates (such as format, overall difficulty, and content area).

Assessing the Importance of Non-Response

In some instances, routine monitoring of non-response will uncover particularly severe problems, either overall or for particular groups of students or items. In such cases, it may be important to analyze the sensitivity of primary NAEP results to those problems. The methods used to scale and analyze NAEP data are unusually complex and arcane, and few consumers of the information produced are likely to be able to judge the likely impact of non-response without access to such sensitivity analyses. To be useful, the results of sensitivity analyses should be reported in all vehicles that include information on non-response, and a description of the analyses should be available, perhaps in the NAEP Technical Reports.

References

Beaton, A. E., & Zwick, R. (1990). ***The effects of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly*** **(Rept. No. 17-TR-21). Princeton, NJ: Educational Testing Service.**

Johnson, E. G. (1988, November). Mathematics data analysis. In A. E. Beaton (Ed.), ***Expanding the new design: The NAEP 1985-86 technical report*** **(pp. 215-242). Princeton, NJ: NAEP/Educational Testing Service.**

Swinton, S. (1991). ***Differential response rates to open-ended and multiple-choice NAEP items by ethnic groups.*** **Unpublished manuscript, Educational Testing Service, Princeton, NJ, October 23.**

Swinton, S. (1992). ***Differential response rates to open-ended multiple-choice items by ethnic group and administration mode—Phase II.*** **Unpublished manuscript, Educational Testing Service, Princeton, NJ, February 3.**

Yamamoto, K., & Jenkins, F. (1992, February). Data analysis for the mathematics assessment. In E. G. Johnson & N. L. Allen (Eds.), ***The NAEP 1990 technical report*** **(Report No. 21-TR-20, pp. 243-276). Princeton, NJ: NAEP/Educational Testing Service.**

Appendix A

Cognitive Complexity Classifications

In addition to examining the ETS learning technique scheme, we developed a cognitive complexity classification scheme to try and capture a dimension we felt may have been the cause of some of the omit and not-reached behavior. The categories of our classification scheme are as follows:

Category 1: Standard classroom problem requiring only a single step or the application of a single concept/procedure/algorithm;

Category 2: Real World problem solving requiring multiple steps (see note on steps below);

Category 3: Standard classroom problem requiring multiple steps; and

Category 4: Novel problem solving. (The novelty implies multiple steps in that the translation of the novelty into familiar terms constitutes a step.)

Trying to distinguish between difficulty and complexity was a continual problem as we attempted to classify the items. Focusing on how a sophisticated mathematician would solve a problem, rather than trying to imagine how a twelfth grader would solve it, helped alleviate this somewhat, but there were still a number of items difficult to classify. Determining the number of steps entailed by a problem was also complex. Here a distinction was made between procedures or algorithms which required multiple steps and a solution process which required multiple steps. In this scheme the former were considered a one-step problem while the latter were considered multistep. This distinction is subtle but is an important aspect of complexity nonetheless. In determining the number of steps, the translation of a real world problem into its mathematical equivalent counts as one step, and then the solution of the appropriate equation counts as another. This reasoning also applies to problems which are presented in a novel or non-standard fashion.

In making the distinction between standard classroom and novel, a judgment was made as to whether or not the problem was likely to have been encountered by an appreciable percentage (ambiguity intended) of the sample enrolled in a "standard" instructional program.