
Voluntary National Tests

in Reading and Mathematics

***EFFECTS OF ITEM SCAFFOLDING ON
STUDENT RESPONSES:
A COGNITIVE LABORATORY STUDY***

October 13, 2000

Prepared by

American Institutes for Research
for the National Assessment Governing Board
in support of contract # RJ97153001

For questions about this document, please contact Christine Paulsen at 978-371-8341

INTRODUCTION

This report describes the findings associated with a special study that is part of the development of the Voluntary National Tests Subtask 2.2.2. During the first and second years of VNT item development, the most common types of item problems found using the cognitive laboratories were ambiguous instructions and unclear language (AIR, 1998; 2000a). In reading, ambiguous instructions were identified as potential item problems in most of the short constructed response items and extended constructed response items. The majority involved instances in which students had difficulty understanding the initial part of the stem in which the question that is specific to the passage is posed. However, in a number of cases, students also had difficulty understanding the concluding instructions, “Give details from the passage” or “Give two details from the passage.” In mathematics, confusion centered around the instructions “Show your work” and “Justify your answer.” All interviewers reported many cases in which students’ written responses did not match their oral responses to specific probes asked by the interviewers. Students’ difficulty in comprehending these basic instructions prevented some students from receiving full credit because the scoring rubrics are often structured around the number and type of details.

These findings point to the need for considering ways to improve the cues that are designed to elicit desired responses from students when responding to constructed response items. It is essential that students understand the requirements for full credit responses. The purpose of this special study was to determine whether there is a way to encourage students to provide full credit responses in written form. There are several methods item writers may use to elicit better responses from students. The method explored in this study is *item scaffolding*. Item scaffolding basically consists of extracting and highlighting the key requirements of an item so that students can clearly see what is expected of them. For example, the following item contains scaffolding (in italics).¹

Example. John drew a four-sided closed figure with four parallel sides and four right angles. Draw a similar figure and label your drawing to show the parallel sides and the right angles.

Remember, to get full credit you must:

- 1 – Draw a four-sided closed figure with four parallel sides and four right angles, and*
- 2 – Label the drawing to show the parallel sides and right angles.*

¹ In order to ensure the VNT item pool remains secure, this is not an actual VNT secure item. No secure VNT items will appear in this report. To see a copy of the items and item protocols used in this study, please contact Christine Paulsen at 202-944-5443. *Only authorized individuals associated with the VNT who have signed an Affidavit of Nondisclosure are given access to secure VNT items.*

The concept behind scaffolding is that, by explicitly restating what is expected from students, items can elicit better responses from students. This is not to imply that the test items are easier or that they contain cues or that they “give away” the answer or reduce content, skill, or knowledge requirements. Instead, scaffolding simply presents information from the item in a more explicit manner to help students better organize their responses.

This report describes the mathematics and reading studies separately. The final section summarizes and synthesizes the findings for both studies, and includes some recommendations for future item development.

MATHEMATICS STUDY

Study Objectives and Design

Our objective was to compare the quality of nine students’ responses to items with scaffolding to the quality of these same students’ responses to items that did not contain scaffolding. Our hypothesis was that students would provide better written responses to items that were scaffolded. We judged the quality of student responses based on the amount of credit received (i.e., no credit, partial credit, or full credit) according to the scoring rubrics provided by item writers.

In addition, we were interested in exploring whether scaffolded items can help students to understand better what is expected from them – are the scaffolded items truly more explicit? To study this, interviewers asked each student the following question after each item, “*How did you know how much work to show for this question, and how did you know when you were finished?*” We analyzed student responses to these questions to see if students would indicate that the scaffolding helped them to know how much work to show and helped them to know when they had responded adequately.

This study used a within-subjects design in which each student served as their own control. For this special study, cognitive laboratories were conducted at AIR offices in Washington, DC and Concord, MA. Nine 8th grade students participated in the mathematics study.

Procedures and Instruments

During the cognitive laboratory interview, the interviewers were guided by an interview protocol. Included in the protocol (for each item) were the item and the correct response, the item type and other descriptive characteristics (stance or strand), general item probes, probes that were

specific to the item, and the expected solution path(s).

Nine 8th grade students responded to the same 10 mathematics items, in the exact same order. The first five constructed response items were from a variety of strands and were not scaffolded. A second set of five items was scaffolded to elicit greater detail from students. The second set of five items was chosen by internal subject matter experts to reflect the same strands as the first five items, and generally the same expected level of difficulty.

Results

Quality of Responses: Earned Credit

Students were awarded credit for their responses to constructed response items based on the scoring rubrics provided by item writers. Cognitive lab interviews scored each of the students' responses, and these scores were checked by senior analyst staff members. Students could earn full credit, partial credit, or no credit. We compared the average credit earned by students on the five control items to the average credit they earned on the five experimental (scaffolded) items.

Of the nine students who participated in the special study, only three appear to have earned slightly more credit on the experimental items. However, note that none of these three students actually completed all five experimental items (see Table 1). Two students earned less credit on the experimental items than the control items.

Table 1:
Average Scores on Control and Experimental Items
(n = number of items, out of 5, actually completed)

Student	Control Items	Experimental Items
1	1.4 (n=5)	1.2 (n=5)
2	1.4 (n=5)	2.0 (n=2)
3	1.8 (n=5)	1.0 (n=5)
4	0.8 (n=5)	0.8 (n=5)
5	0.4 (n=5)	0.4 (n=5)
6	1.6 (n=5)	1.6 (n=5)
7	1.4 (n=5)	2.0 (n=3)
8	1.4 (n=5)	1.3 (n=3)
9	1.4 (n=5)	1.5 (n=2)

Knowing What is Expected

For each item, interviewers asked students to indicate how they knew when they were done answering the question. We wanted to study the extent to which scaffolding might help students better understand what is expected from them in responding to mathematics items. On control items, 68% of the time students said they wrote as much as they could think of to respond to the item or they simply guessed that they were done – they said there was no information in the item itself that helped them determine how much work to show. Thirty-two percent of the time on control items, students said the instructions embedded within the item helped them to know how much information to provide to receive full credit.

On experimental items, students guessed that they had written enough only 18% of the time. Eighty-two percent of the time, students indicated that the item requirements outlined in the scaffolding helped them to determine whether they had responded sufficiently.

Table 2:
Reasons Students Gave for Knowing when they had
Responded Sufficiently to Control and Experimental Items

Reasons Students Gave for Knowing when they had Responded Sufficiently	Percentage of Times Students Used this Reason	
	Control Items	Experimental Items
Guessed, hunch, just wrote until they couldn't think of anything else to include	76%	16%
Based on requirements specifically outlined in the item	24%	84%

Anecdotal Evidence

Overall, students expressed a definite preference for the scaffolded items. Different students made the following comments:

- *It is clearest when there is a list of requirements that I need to do to get full credit.*
- *I like the way the parts were listed in item #11 (an experimental item) as opposed to questions that just asked you to show your work as in item #1 (a control item).*
- *When they had the list, it was best.*
- *The requirement things are helpful because they tell you exactly what to do.*
- *The ones that had the listed steps were more clear and better.*

READING STUDY

Study Objectives and Design

Our objective for the reading study was to compare the length and quality of 4th grade students' responses to items with scaffolding to the quality of these same students' responses to items that did not contain scaffolding. Our basic research questions were: (1) Does scaffolding lead to longer responses, and if so, (2) are these longer responses of higher quality?

To judge the length of student responses we simply counted the number of words written by each student for each item. We judged the quality of student responses based on the amount of credit received (i.e., no credit, partial credit, or full credit) according to the scoring rubrics provided by item writers.

In addition to questions of length and quality, we were also interested in exploring whether scaffolded items helped students to better understand what is expected from them – are the scaffolded items truly more explicit? As in the mathematics study, interviewers asked each student the following question after each item, “*How did you know how much to write for this question, and how did you know when you were finished?*” We analyzed student responses to these questions to see if students would indicate that the scaffolding helped them to know how much to write and helped them to know when they had responded adequately.

Data from eight 4th grade students were available for this study. An additional ninth student was interviewed, but did not finish the interview in time to respond to any of the experimental items. Therefore, this student's data was not included in the results reported here.

Procedures and Instruments

Eight students responded to the same six reading items presented in the same order to each student. The first passage contained three constructed response items that were not scaffolded (one ECR and two SCRs). The second passage contained three items that *were* scaffolded (again, one ECR and two SCRs). Internal experts attempted to choose passages and items that matched well in terms of content, type, and difficulty.

Results

Response Length

We counted the actual number of words in each student response. Since there were two SCRs in the control items and two SCRs in the experimental items, these word counts were averaged across both items. The actual ECR and average SCR word counts are summarized in Table 3 below:

Table 3:
Word Counts for Control and Experimental Items

Student	Control Items		Experimental Items	
	SCRs (n=2)	ECR (n=1)	SCRs (n=2)	ECR (n=1)
1	22.5	22	21.5	44
2	38.0	33	30.0	51
3	34.5	48	32.5	42
4	39.5	47	35.0	98
5	88.0	101	29.0	120
6	68.0	79	35.5	75
7	48.5	54	40.0	33
8	42.5	51	22.0	72

Five out of the eight students wrote longer responses for the experimental ECR item than for the control ECR. However, with only two ECR items for comparison, it is impossible to say whether the difference is due to the scaffolding or differences in item difficulty or some other factor. Conversely, for SCR items, scaffolding did not appear to elicit longer responses from students. Again, while we attempted to choose comparable items, it is difficult to say whether these differences are due to scaffolding or differences in item difficulty.

Response Quality: Earned Credit

Data on the length of student responses is made more meaningful when combined with data on actual student performance. Thus the second research question was: If student responses are longer for experimental items, are they also getting more credit for these responses than they did on control items? Table 4 summarizes the amount of credit earned by each student.

Table 4:
Credit Earned on Control and Experimental Items

Student	Control Items		Experimental Items	
	SCRs (n=2)	ECR (n=1)	SCRs (n=2)	ECR (n=1)
1	1.5	1	1	1
2	1.5	1	0.5	1
3	1	1	1	1
4	1.5	1	0.5	1
5	2	2	1	1
6	1.5	2	1	1
7	1	2	2	1
8	1	2	2	1

With respect to SCR items, only two students did better on the experimental items than they did on the control items. Five out of the eight students actually did better on the control items than they did on the experimental items. This may be due to an unexpected difference in item or passage difficulty. With respect to the ECR items, half of the students received more credit on the experimental item than the control item, while half fared worse on the experimental item. Of the five students who actually wrote longer responses to the experimental items, only two of those students received more credit for what they wrote.

Knowing What is Expected

For each item, interviewers also asked students to indicate how they knew when they were done responding to each requirement of the item. As indicated in Table 5, on control items, 54% of the time students said they simply guessed that they had written enough or they wrote as much as they could think of in order to respond to the item. Forty-six percent of the time on control items, students said the instructions embedded within the item helped them to know how much information to provide to receive full credit.

On experimental items, students guessed only 38% of the time. Sixty-three percent of the time, students indicated that the item requirements outlined in the scaffolding helped them to determine whether they had responded sufficiently.

Table 5:
Reasons Students Gave for Knowing when they had
Responded Sufficiently to Control and Experimental Items

Reasons Students Gave for Knowing when they had Responded Sufficiently	Percentage of Times Students Used this Reason	
	Control Items	Experimental Items
Guessed, hunch, just wrote until they couldn't think of anything else to include	54%	38%
Based on requirements specifically outlined in the item	46%	63%

Anecdotal Evidence

As was the case with mathematics items, most students expressed a preference for the scaffolded items. Different students made the following comments:

- *The (scaffolded format) helps you. It asks the question and then you write an answer.*
- *The second kind, where the questions were split up, was easier than the first.*
- *It was easier when they separated the questions and gave two spots to answer.*
- *It was easier to write it down that way when they broke the question up rather than writing in one big space.*
- *It was easier to answer the question when it was broken down into two separate parts. It was easier to think about it.*

However, one student stated that the item format did not matter to her – neither was easier to understand. A second student also said she did not have a preference. She could “*see how the second section could be easier for most kids*” but she did not find it easier or more difficult.

CONCLUSIONS AND IMPLICATIONS

These results suggest that scaffolding can help students understand better what is expected from them when responding to mathematics and reading items. On the experimental items, students said they referred back to the scaffolding to make sure they had addressed each of the item

requirements, rather than having to guess about whether they met all the item requirements. Students more clearly understood how much work to show to earn full credit, and preferred the scaffolded item format.

However, just because students knew what was required of them, and were more clear about how much of a response to provide, results do not support the hypothesis that the item scaffolding lead to better quality responses. While the results do not strongly support the hypothesis that scaffolding items leads to higher quality responses, the small sample size and missing data make it impossible to conclude that scaffolding does not elicit more information from students. Moreover, while internal subject matter experts attempted to match items in terms of type and difficulty, it is possible that unexpected differences existed between control and experimental items and that these differences are a confounding factor.

This exploratory study does suggest that students prefer scaffolded items, and that such items help students to make decisions about whether they have responded in a sufficient manner to earn full credit. The item scaffolding may not directly lead to better mathematics or reading scores, but it can help students remember or otherwise be clear about all the steps they need to take to earn full credit, and also helps them to better organize their responses.

Additional studies of the impact of item scaffolding on student responses should be conducted with larger sample sizes to confirm or expand these findings. However, this study does suggest that students can benefit from techniques like item scaffolding. It can help students avoid wasting precious testing time worrying about whether they have addressed all the necessary parts of an item – leaving more time for thinking constructively and processing substantive information directly related to the test items themselves. While the impact of scaffolding on students with learning disabilities and limited English proficiency has not been explored in this study, other cognitive lab studies conducted this year seem to indicate that such students would also benefit from techniques such as scaffolding (AIR, 2000b). This small study suggests that item scaffolding may help to remove a potential source of measurement error, and for that reason it is a technique that should be considered by test developers.

REFERENCES

- American Institutes for Research. (1998). *Cognitive Laboratory Report: Year 1*. Washington, DC: Author.
- American Institutes for Research. (2000a). *Cognitive Laboratory Report: Year 2*. Washington, DC: Author.
- American Institutes for Research. (2000b). *A Cognitive Laboratory Investigation of the Performance of Learning Disabled Students on VNT Items*. Washington, DC: Author.