

# **What Types of Survey Items Can Elicit Valid Responses from Fourth and Eighth Grade Students?**

**30 September 2002**

**Roger Levine and Mette Huberman**  
**American Institutes for Research, Palo Alto, CA**  
**Arnold Goldstein, Project Officer**  
**U.S. Department of Education, National Center for Education Statistics**

## **Introduction**

The National Assessment of Educational Progress (NAEP) is involved in a major effort to measure the educational achievement of our nation's children. As part of this effort, when students take achievement tests, they answer questions about a variety of background factors that are known or believed to be related to performance on these tests. Some of the questions are related to home background factors (e.g., parent education, TV watching, and homework). Other questions ask about instructional background factors, such as the use of computers in math or the prevalence of group activities in reading. Teachers also answer instructional background questions, including how they teach math or reading.

There has been concern about the quality of the student background measures. Data from fourth graders on specific factors (e.g., parent education) were seen as not being reliable due to high omission rates. For similar reasons, some data from eighth grade students were also of questionable reliability, in particular for minority students. Accordingly, a study was undertaken in 1996–97 to improve the quality of the fourth and eighth grade home background items through the use of cognitive survey methodologies. A cognitive interviewing protocol was developed, which included a validation component. Parents of the students were interviewed, providing validation information. This information enabled identification of item problems that would not otherwise have been detected.

In 1997–98, the study was extended to investigate instructional background items completed by students and their teachers. Again, a cognitive interviewing protocol was developed, which included a validation component. However, this time the students' teachers provided the validation information. This paper summarizes the findings from both parts of this study. A total of 137 students (83 fourth graders and 54 eighth graders), 71 parents, and 12 teachers participated in the study.

The major foci of this paper are the types of items (behavioral frequency and time estimation items) that were found to be particularly difficult for children to answer and comprehension issues associated with the administration of surveys to children. In addition, a commonly used survey format was found to be a source of response error for both children and adult respondents.

## **Background**

Cognitive interviewing (also known as verbal reporting [Willis et al., 1991]), of questionnaire respondents is a form of interview used to uncover the mental processes involved when a respondent reads and responds to survey questions (Willis et al., 1999). Thus, cognitive interviews, as a form of survey pretesting, are effective in determining how respondents comprehend survey items and what strategies they use to devise answers. Such interviews are primarily conducted to identify sources of respondent confusion and misunderstanding (Krosnick, 1999; Fowler and Cannell in Schwarz and Sudman, 1996; Schaeffer and Maynard in Schwarz and Sudman, 1996). More specifically, they may lead to the verification of an expected question problem or the discovery of one that was unanticipated (Willis et al., 1999; DeMaio and Rothgeb in Schwarz and Sudman, 1996.) Cognitive interviewing can facilitate not only the finding of a problem but also the fixing of a problem (Willis et al., 1999).

In most cognitive interviews, two basic techniques are utilized: “think-alouds” and verbal probing (Willis et al., 1999). In the think-aloud technique, the interviewer asks the subject to report what he or she is thinking as he or she is answering a question. Often a simple directive such as “Can you tell me what you’re thinking?” is used. The interviewer records the process the respondent engages in as he or she arrives at an answer. In addition, specific probes related to each item (e.g., paraphrasing of the question or requests for definitions of words and phrases) are developed and administered after the participant has produced a response to the survey question.

The use of validating information can be of tremendous benefit to the cognitive interviewer. Although think-alouds and normal probing enable the identification of many item problems, there are occasions when an individual’s think-aloud and responses to probes will fail to reveal the existence of an incorrect response. But, if the interviewer knows the response is problematic (as a result of validating information), further probing can be employed until reason(s) for the otherwise undetectable item problem are determined. Protocols employing validation data have been successfully employed, enabling the modification of self-administered survey items. These validation data also enabled assessments of the impacts of item modifications. Accordingly, it was possible to demonstrate that the use of validating information (provided by a parent or guardian) in cognitive investigations of survey items completed by children lead to the development of revised items with lower error rates than their unmodified counterparts (Levine et al., 2001). Validation data also lend themselves to simple, tabular presentations of the effectiveness of an item.

In the studies reported upon, parents were the source of validating information about their children’s home behaviors and teachers were the source of validating information about the school and school-related behaviors of their students. In addition, teacher self-report behavior was validated through a calendar exercise. Effective procedures for validating self-report data employing focused retrieval, varied retrieval, and attempts to evoke multiple representations of the construct of interest through the reconstruction of a calendar/diary for a time period of interest have been used to validate self-reports of hours worked (Edwards, Levine, and Cohany, 1989). In a similar fashion, in order to improve the quality of eyewitness reports of a crime, cognitive science techniques have been employed in a procedure called (coincidentally) “cognitive interviewing.” The principles underlying this type of cognitive interviewing are (Fisher and Quigley, 1992):

- 1) *Context reinstatement.* The same psychological environment in which the event occurred is created to increase the validity of recall. (Tulving and Thomson, 1973)
- 2) *Focused retrieval.* The interviewer gets the respondent to expend effort and engage in uninterrupted concentration. (Kahneman, 1973)
- 3) *Extensive retrieval.* The more retrieval efforts the respondent makes, the more successful recall will be. Thus, the respondent is encouraged to search through memory even if she thinks that she has recalled everything. (Roediger and Payne, 1982)
- 4) *Varied retrieval.* The use of different retrieval probes is more effective than the use of a single retrieval probe. (Anderson and Pichert, 1978)
- 5) *Multiple representations.* The construct of interest will have different mental representations in the respondent’s memory. Each different representation is evoked and probed separately. (Fisher and Chandler, 1984)

These procedures have also been shown to be effective in enhancing dietary recall (Fisher and Quigley, 1992).

## **Methodology**

*Survey and Protocol Development.* The survey items studied were taken from large, U.S. Department of Education studies, including the National Assessment of Educational Progress (NAEP) background questionnaires and the National Educational Longitudinal Study (NELS:88).

Selected items were used to create fourth and eighth grade student surveys. For the study of home background factors, about half of the items were selected randomly. The other half was purposively selected because they were known or felt to be problematic. For the study of instructional background factors, all of the items selected were known or believed to be problematic. It must be emphasized that these are not a random selection of NAEP items.

The home background items were used to create surveys that were administered to fourth graders in the first wave of the study. Since only about a dozen home background items could be studied in a session, the 25 items were divided into surveys comprised of 12 and 13 items. Twenty-five fourth graders participated in the first wave. After the first wave of the study, data were analyzed and used to inform item revision. Revised versions of the items were administered to a second cohort of 23 fourth graders and a group of 23 eighth graders

For the instructional background items, fourth grade student surveys were created in math, science, and reading, and eighth grade student surveys in math and science. The three fourth grade subject area surveys consisted of between 12 to 19 items. Since three subject area surveys (with their associated protocols) could not be administered in a two-hour cognitive interview session, each fourth grader was administered two of three surveys—mathematics and science, mathematics and reading, or science and reading. Because of the departmentalized nature of middle schools, as mentioned earlier, eighth graders answered either a 23-item mathematics survey or a science survey with 23 items. Thirty-five fourth graders and 31 eighth graders were engaged in the cognitive testing of instructional background items.

Protocols related to each survey were developed. The protocols provided a variety of optimal probes (e.g., word definition and paraphrasing), to be used when deemed appropriate. In addition, interviewers were trained to create unique probes during the interviews tailored to the individual respondents' interpretations of the items.

*Participant Recruitment and Selection.* In order to get a heterogeneous group of participants, several different school districts were contacted in the San Francisco Bay Area. As an incentive for participation, students (and their parents) were offered \$50 and teachers \$100 for their time. Potential participants were screened to allow selection of a diverse sample with respect to the household's annual income level and the race/ethnicity of the student. A total of 137 students (83 fourth graders and 54 eighth graders), 71 parents, and 12 teachers (6 fourth grade teachers and 6 eighth grade teachers) participated in the study.

### **Interview Procedures**

*Parent Interviews.* While their child was answering the survey questions related to home background factors, parents were interviewed in a different room. They were asked how they thought their child would respond to each item, what the correct answer was, and, if these responses were different, why there was a discrepancy. Specific questions about each item were also asked, to provide essential contextual information.

After the interview with the parent, a videotape of the child's interview was played back for viewing with the child and the member of the research team who had interviewed the parent. During this playback, the researcher (who previously interviewed the parent) was aware of the "correct" response. Whenever the child's responses were discrepant from the parent's report, the tape was stopped and the reasons for these discrepancies were probed.

Concurrently, the parent (in an adjacent room, behind a one-way mirror) was observing the child-researcher interaction. When the child's responses were discrepant, the parent was asked for possible

explanations. Parents could (and did) inform about the veracity of the child's responses thereby permitting the detection of confabulation.

*Teacher Interviews.* The teacher interviews lasted about two hours and consisted of two phases. In the first phase, the survey items were administered. The teacher was asked to read the questions aloud and was reminded to think aloud to provide insights into the cognitive processes that he or she employed in responding to the items. In addition, specific probes and paraphrasing requests were used to further inform about the response process. Typical probes included: "What do you think they mean by [technical term]?" and "What do you think this question is asking?"

The second phase of the interview was designed to validate some of the teachers' responses in Phase 1. For context reinstatement purposes, interviews were conducted in the teacher's classroom. The teacher was asked to reconstruct the past week's activities in class through a calendar exercise. To this end, weekly matrices were designed to help the teacher recall the frequency of certain instructional practices occurring each day. The teacher recalled each day of the week by first thinking of important or atypical events that occurred during the past week (e.g., staff meetings, sick students, and special events). These were written onto the matrix to serve as cognitive anchors and facilitate recall of each day's events. Then the teacher was asked about the day's lesson in the subject area of interest (e.g., what the teacher taught that day, whether he or she used any special materials such as manipulatives, and whether the teacher utilized technology). To evoke multiple representations, the teacher was also asked to think about a specific student and what this student did during the lesson in question. All of these extensive retrieval activities were intended to facilitate recall. After the interviewer felt these efforts resulted in retrieval of a clear representation of the day's lesson, the teacher was then asked to estimate, which of several instructional practices occurred on that day in the subject area of interest. This process was repeated for each subject and each day. Responses were documented on the matrix, as well. Then, adjustments were made to account for the typicality of the week, and these frequencies were then compared to the frequencies given in the first phase of the interview. Discrepancies were probed, and the most accurate answer (as determined by the teacher) was noted.

A summary of each interview focusing on item problems was prepared subsequent to each teacher interview. Before the student interviews were conducted, teacher responses and any other relevant information from the teacher interviews were recorded on the student protocols to facilitate the identification of discrepancies and the triggering of probes.

*Student Interviews.* The students were asked to read the questions aloud. This facilitated detection of potential language and comprehension problems. For example, when a student could not read or pronounce a word, it was an indication of a comprehension problem. In these cases, the interviewer would make sure to probe the student's understanding of the particular word.

The students were continually encouraged to think aloud. Probing and paraphrasing requests were utilized to inform about the student item response process. When a student's response varied from that of their teacher or parent, the interviewer tried to determine the reason for the discrepancy by administering additional probes about the item (e.g., asking for further elucidation about how frequency estimates were produced or verifying comprehension of the item).

After each student interview, a summary was prepared focusing on item problems and the reasons for any discrepancies between the students' and the teachers' or parents' answers.

## Analysis

In order to summarize results, students' responses were compared to the teachers' and parents' responses. Discrepancy rates for items were calculated by comparing the number of mismatches between the students' and their teachers' or parents' responses. These discrepancy rates were calculated across all student-teacher and child-parent item pairs. When discrepancies occurred, the student and teacher/parent summaries were analyzed to identify the reasons for the discrepancies. For most items, the adult's response was considered to be the correct response. However, situations would occasionally arise which indicated that the adult had misinterpreted the question or that the student was more accurate. In these cases, the parent's "corrected" response was used for validation.

Teacher's validated responses to behavioral frequency items were compared with their initial survey responses.

## Results and Discussion

From the think-alouds and the validation data provided by teachers and parents, it was possible to compare and validate the students' answers against the teachers' and parents' responses, identify inconsistencies, and determine the reasons that these inconsistencies occurred. Four general areas of item problems emerged: 1) Behavioral frequency items; 2) Time frame and time estimation; 3) Language and comprehension; and 4) Use of list formats. These issues and recommendations for item revisions are discussed below.

*Behavioral Frequency Issues: Home Background Items.* Home background items were categorized as behavioral frequency items, time estimation items, and other items. Error rates for these items are summarized below. Error rates for the initial administration of survey items to fourth graders are summarized in the "4th Grade (1)" column; error rates for the administration of the revised items to a different group of fourth graders and a group of eighth graders are summarized in the "4th Grade (2)" and "8th Grade (2)" columns.

**Summary of Error Rates for Fourth and Eighth Graders**

| Type of Item         | 4th Grade (1) | 4th Grade (2) | 8th Grade (2) |
|----------------------|---------------|---------------|---------------|
| Behavioral frequency | 64.8%         | 52.8%         | 32.7%         |
| Time estimation      | 41.7%         | 42.3%         | 42.3%         |
| Other                | 33.2%         | 20.3%         | 20.2%         |
| Total                | 43.0%         | 32.2%         | 26.6%         |

Behavioral frequency items required the estimation of a rate. They all began with the phrase, "How often do(es)..." These items require the possession and application of rate estimation skills -- skills that many fourth graders do not possess. Although item rewordings can help clarify the construct being addressed by the item, they cannot overcome a lack of rate estimation skills. The suggested rewordings for this type of item resulted in a lower error rate for fourth graders (52.8 percent vs. 64.8 percent). Eighth graders seem to possess better rate estimation skills: their error rate on these items was 32.7 percent.

Other items refer to items that did not require time estimation, rate estimation, or averaging skills. Suggested modifications for these items, based on analysis of the cognitive interviewing data, resulted in substantially lower error rates for fourth graders (20.3 percent vs. 33.2 percent). Fourth graders in the

initial and subsequent wave differed in a number of ways that might influence the validity of their responses. The first phase was conducted in the spring of the school year; the second, in the fall of the next school year. According, fourth graders in the second wave were almost certainly younger than students in the first wave. Furthermore, there was a preferential selection of minority students and lower SES students in the second wave. These differences would generally be expected to operate to the disadvantage of students in the second. Nonetheless, their error rate (to the revised survey items) was lower. This strongly suggests that the cognitive interviews informed successful revision of this type of item.

*Behavioral Frequency Issues: Instructional Background Items.* The estimates produced by teachers to behavioral frequency items after the validation exercise were compared to their responses to the original presentation of these items. These items were rated on a four-point frequency scale: (1) Never or hardly ever, (2) 1-2 times per month, (3) 1-2 times per week, and (4) Almost every day. The proportion of responses that changed as a result of the enhanced recall procedures and the proportion of responses that changed by two or more scale points (defined as a major discrepancy) are summarized below.

**Discrepancy Rates of Fourth Grade Teachers to Selected Behavioral Frequency Items, by Subject Area**

| Subject Area | Discrepancy Rate |                   |
|--------------|------------------|-------------------|
|              | Any Discrepancy  | Major Discrepancy |
| Science      | 37%              | 2%                |
| Math         | 36%              | 11%               |
| Reading      | 36%              | 3%                |

NOTE: Results are based on the responses of six teachers.

**Discrepancy Rates of Eighth Grade Science and Mathematics Teachers to Selected Behavioral Frequency Items, by Subject Area**

| Subject Area | Discrepancy Rate |                   |
|--------------|------------------|-------------------|
|              | Any Discrepancy  | Major Discrepancy |
| Science      | 12%              | 0%                |
| Math         | 25%              | 8%                |

NOTE: Results are based on the responses of three science and three mathematics teachers.

Major discrepancies were relatively rare. They generally were due to the teacher’s misinterpretation of the item’s intent. Minor discrepancies reflected both the difficulty of the task (producing behavioral frequency estimates) and gaps between the scale points. Gaps occurred between all of the scale points. Behaviors performed about three times per week could either be classified as “Almost every day” or “1-2 times per week;” behaviors performed approximately three times a month could be classified as “1-2 times per week” or “1-2 times per month;” behaviors that occurred between three and six times per year could be classified as “Never or hardly ever” or “1-2 times per month.” Minor changes in a behavioral frequency estimate would result in its classification in an adjacent response category.

Discrepancy rates for behavioral frequency items for students were also determined. For several items, such as items asking about a student’s participation in classroom discussions, variations in individual student behavior were anticipated. Some students may talk to the class about their mathematics work more than others. However, many of these behaviors are ones for which the between student variation

within a classroom would be expected to be minimal. Taking mathematics tests is one such example; doing problems from textbooks or worksheets are other examples.

### Discrepancy Rates for Fourth Grade Science Items

|  |               |
|--|---------------|
| <b>When you study science in school, how often do you do each of the following:</b>            | <b>(n=16)</b> |
| Discuss science in the news  | 69%           |
| Do hands-on activities in science  | 75%           |
| Talk about measurements and results from your hands-on activities                              | 75%           |
| Use a computer for science   | 13%           |
| Use library resources for science  | 69%           |
| <b>When you study science in school, how often does your teacher do each of the following?</b> |               |
| Talk to the class about science  | 81%           |
| Do a science demonstration   | 88%           |
| Show a science videotape or science television program   | 63%           |
| Use computers for science (e.g., such as science software, telecommunications)                 | 13%           |

NOTE: The number of students (n) is the modal number of respondents to each item.

### Discrepancy Rates for Fourth Grade Mathematics Items

|   |               |
|---|---------------|
| <b>When you do mathematics in school, how often do you do each of the following:</b>  | <b>(n=19)</b> |
| Do mathematics problems from textbooks  | 68%           |
| Do mathematics problems on worksheets   | 74%           |
| Solve mathematics problems with a partner or in small groups  | 74%           |
| Work with objects like rulers, counting blocks, or geometric shapes   | 74%           |
| Write a few sentences about how you solved a mathematics problem  | 61%           |
| Take mathematics tests  | 32%           |
| Talk to the class about your mathematics work   | 89%           |
| Do 10 or more practice problems in mathematics by yourself  | 79%           |
| Discuss solutions to mathematics problems with other students   | 58%           |
| Use a computer  | 37%           |
| Use a calculator  | 42%           |
| <b>This year in school, how often have you taken mathematics tests where you were asked to provide detailed solutions to problems you had not worked on before?</b> | <b>88%</b>    |

NOTE: The number of students (n) is the modal number of respondents to each item.

### Discrepancy Rates for Fourth Grade Reading Items

| <b>When you have reading assignments in school, how often does your teacher do each of the following:</b>                            | <b>(n=16)</b> |
|--|---------------|
| Ask you to do a group activity or project about what you have read   | 50%           |
| Ask you to talk to other students about what you have read   | 79%           |
| Ask you to write about something you have read   | 33%           |
| Help you break words into parts  | 62%           |
| Help you understand new words  | 60%           |
| This year in school, how often have you been asked to write long answers to questions on tests or assignments that involved reading? | 67%           |

NOTE: The number of students (n) is the modal number of respondents to each item.

For fourth graders, the average discrepancy rate for these items was 62 percent. Alternatively, these rates can be presented as their complement: agreement rates. The average agreement rate was 38 percent. Through random guessing, agreement rates for 4-point scale items of 25 percent.

Discrepancy rates for eighth grade science and mathematics items are presented below.

### Discrepancy Rates for Eighth Grade Science Items

| <b>When you study science in school, how often do you do each of the following:</b>            | <b>(n=11)</b> |
|--|---------------|
| Discuss science in the news  | 18%           |
| Work with other students on a science activity or project                                      | 82%           |
| Give an oral science report  | 55%           |
| Do hands-on activities or investigations in science  | 73%           |
| Talk about the measurements and results from your hands-on activities or investigations        | 82%           |
| Go outside to observe or measure things  | 36%           |
| Design and carry out your own science investigations   | 55%           |
| <b>When you study science in school, how often does your teacher do each of the following:</b> |               |
| Talk to the class about science  | 27%           |
| Do a science demonstration   | 45%           |
| Use computers for science (e.g., science software, telecommunications)                         | 18%           |
| About how often does your science class go on a science field trip?                            | 18%           |
| About how often does a guest speaker come to speak to your science class?                      | 9%            |
| Which best describes the science course you are taking?  | 67%           |
| About how often do you study science in school?  | 36%           |
| Do either you or your teacher save your science work in a portfolio?                           | 55%           |
| Do you ever do science projects in school that take a week or more?                            | 27%           |

NOTE: The number of students (n) is the modal number of respondents to each item.

### Discrepancy Rates for Eighth Grade Mathematics Items

| <b>When you do mathematics in school, how often do you do each of the following:</b> | <b>(n=20)</b> |
|--|---------------|
| Do mathematics problems from textbooks   | 65%           |
| Do mathematics problems on worksheets  | 60%           |
| Solve mathematics problems with a partner or in small groups                         | 55%           |
| Work with measuring instruments or geometric solids                                  | 55%           |
| Write a few sentences about how you solved a mathematics problem                     | 50%           |
| Take mathematics tests   | 25%           |
| Talk to the class about your mathematics work  | 80%           |
| Do 10 or more practice problems by yourself  | 38%           |
| Discuss solutions to mathematics problems with other students                        | 60%           |
| Use a computer   | 45%           |
| Use a calculator   | 20%           |
| Write reports or do mathematics projects   | 20%           |
| Work and discuss mathematics problems that reflect real-life situations              | 65%           |
| Do either you or your teacher have a portfolio with your mathematics work in it?     | 30%           |
| What kind of mathematics class are you taking this year?                             | 23%           |

NOTE: The number of students (n) is the modal number of respondents to each item.

The average discrepancy rate for these items administered to eighth graders was 49 percent. The agreement rate, the complement of this percentage, was 51 percent. This agreement rate was greater than that of fourth graders (38 percent) for comparable items.

The major reasons for the high discrepancy rates associated with behavioral frequency items were:

- 1) Many of the behaviors of concern were not being interpreted as intended by the item writers.
- 2) Many individuals (especially students) lack the cognitive abilities to synthesize a behavioral frequency accurately, particularly when:
  - ◆ the behavior does not occur on a regular basis
  - ◆ the behavior occurs frequently,
  - ◆ the behavior is of low salience, and
  - ◆ the time period (i.e., the denominator for rate calculations) is either ambiguous, unspecified, or long.

Behavioral frequency items have been extensively studied by survey researchers (Sudman, Bradburn, and Schwarz, 1995). Respondents typically estimate rare behaviors or events by counting them. However, common events or behaviors, which are the focus of many of the behavioral frequency questions, are not distinct enough for counting. So, estimation strategies have to be employed. The estimation strategies used are a function of the cognitive skills, abilities, and motivation of the respondent. There is little evidence that the fourth and eighth graders studied possess the skills and abilities required to accurately estimate frequencies for common, low salience behaviors. Even when students understood what sorts of behaviors they were being asked about and could recall specific instances of their occurrence, they were often unable to estimate their frequency of occurrence accurately. Guessing was a common strategy.

*Time Frame and Time Estimation.* Several of the home background items were labeled as “time estimation” items. These items, which asked about time spent reading for fun, doing homework, and watching television, require the abilities to estimate time duration and “average” these durations. These appear to be cognitively challenging tasks for both fourth and eighth graders. Fourth graders have difficulty estimating unstructured time. Some fourth graders also include total time; that is, breaks and other activities engaged in during “homework time” are included in their estimates.

As with the proposed behavioral frequency item modifications, item rewordings may help lower error rates through clarification of the constructs being addressed. However, rewordings cannot teach the time estimation and averaging skills required for the production of correct answers. The suggested item modifications for this type of item resulted in a slightly higher error rate for fourth graders (42.3 percent vs. 41.7 percent). This error rate was the same as the error rate for eighth graders (42.3 percent).

*Language and Comprehension.* The first stage of the item response process is comprehension and interpretation of the item (Tourangeau, 1984; Jobe and Mingay, 1989; Willis et al., 1991). If failure occurs at this stage – that is, if the respondent does not understand what is being asked, there is a strong chance of an inaccurate response. Several language and comprehension problems were found with both fourth and eight grade students.

Many fourth graders had trouble understanding or interpreting the following words and phrases as intended by the item writers:

- |                    |                         |                       |
|--------------------|-------------------------|-----------------------|
| - undecided        | - telecommunications    | - talking about       |
| - e.g.             | - science demonstration | measurements and      |
| - novels           | - practice problems     | results               |
| - geometric shapes | - discussing solutions  | - breaking words into |
| - science software |                         | parts                 |

Replacements such as “not sure” instead of “undecided,” “for example” instead of “e.g.,” and “books with chapters” instead of “novels” seem to work better with students at this grade level. However, fourth grade students generally can not understand technical words and long phrases, so these should be avoided in fourth grade questionnaires. Eighth grade students also had trouble understanding terms such as: “integrated or sequential math,” “applied mathematics (technical preparation),” “geometric solids,” and “hands-on activities or investigations.” Thus, similar to the fourth graders, these types of words and phrases should be minimized with eighth graders.

A surprising comprehension problem occurred in the following home background item where one quarter (25 percent) of the fourth graders to whom the item was administered answered incorrectly:

Does either your father or your stepfather live at home with you?

Yes                         No

It was determined that some children answered “no” to the question because of the conditional “or”. Since their parents were not divorced, they did not have stepfathers. It was therefore impossible for a stepfather to be living at home with them, so they felt the answer to this question was “No.” Other children answered incorrectly, explaining that their fathers were not really living with them, since they worked almost all of the time. As a result of this information, this item was decomposed into two questions (i.e.,

“Does your father live at home with you?” “Does your stepfather live at home with you?”) The modified version was retested and discrepancy rates were reduced to negligible levels.

If a technical term or construct is followed by examples, children often fail to generalize the construct and respond only to the specific examples mentioned in the question. An example of this is the following fourth grade science item (shown in bold).

|  |                          |
|--|--------------------------|
| Have you ever done hands-on activities or projects in school with any of the following? Fill in <b>all</b> boxes that apply. |                          |
| 2. Electricity (for example, batteries and flashlight bulbs)   | <input type="checkbox"/> |
| <b>3. Chemicals (for example, mixing or dissolving sugar or salt in water)</b>   | <input type="checkbox"/> |
| 4. Simple machines (for example, pulleys and levers)   | <input type="checkbox"/> |

This item had a 70% discrepancy rate. As shown below, 16 out of 23 students answered the question inaccurately.

| Teacher Responses | Student Responses |    |
|-------------------|-------------------|----|
|                   | Yes               | No |
| Yes               | 1                 | 0  |
| No                | 16                | 6  |

Fourth grade students overreported hands-on activities or projects with chemicals. Many students did not understand the meaning of the word “chemicals” and focused on the examples provided. They included any situation that involved mixing or dissolving sugar or salt in water (e.g., baking a cake, making lemonade, and doing an experiment with popcorn).

*Use of List Formats.* Items that were presented in a list format (e.g., How often do you do each of the following?) produced problems because of lost context. That is, the respondents often forgot the stem and responded to the items as stand-alone items. This problem was not restricted to fourth graders. For instance, 5 out of 20 eighth graders lost context when they were asked the following math item: “When you do mathematics in school, how often do you do each of the following? Use a computer.” This item was number 10 in a list (as shown below).

|   |                          |                             |                              |                             |
|---|--------------------------|-----------------------------|------------------------------|-----------------------------|
| When you do mathematics in school, how often do you do each of the following? Fill in one box on each line. |                          |                             |                              |                             |
|   | <b>Almost every day</b>  | <b>Once or twice a week</b> | <b>Once or twice a month</b> | <b>Never or hardly ever</b> |
| 1. Do mathematics problems from textbooks   | <input type="checkbox"/> | <input type="checkbox"/>    | <input type="checkbox"/>     | <input type="checkbox"/>    |
| 2. Do mathematics problems on worksheets  | <input type="checkbox"/> | <input type="checkbox"/>    | <input type="checkbox"/>     | <input type="checkbox"/>    |
| .   |                          |                             |                              |                             |
| .   |                          |                             |                              |                             |
| 9. Discuss solutions to mathematics problems with other students  | <input type="checkbox"/> | <input type="checkbox"/>    | <input type="checkbox"/>     | <input type="checkbox"/>    |
| 10. Use a computer  | <input type="checkbox"/> | <input type="checkbox"/>    | <input type="checkbox"/>     | <input type="checkbox"/>    |

When the five students responded to item 10, they had lost the math context of the question. As a result, they overreported their computer usage by answering about their use of a computer anywhere and for any purpose. Maintaining context through redundancy in the list can eliminate item problems like this. So, it was recommended to change item 10 to “Use a computer *for mathematics in school*.”

Teachers, too, lost context, resulting in erroneous responses. After this loss of context was observed, we realized that it might be an artifact of the protocol. That is, thinking aloud after each item might distract the respondent and be responsible for their forgetting the content of the item stem. So, our protocol was modified for list items. Prior to the item, respondents were instructed to answer all of the questions in the list and then tell us what they were thinking as they answered each item. Even with this revision, several instances of lost context were observed.

## **Conclusions**

From the interviews with students and the validation data provided by their teachers and parents, it was possible to identify discrepancies between the students and the teachers’ and parents’ responses, and the reasons that these discrepancies occurred. Several different types of item problems were detected through this process. For example, numerous problems were found in fourth and eighth graders’ ability to understand long and technical words and phrases such as “geometric shapes,” “science demonstration,” “integrated or sequential math,” and “hands-on activities or investigations.” Words and phrases like these should be avoided in fourth and eighth grade questionnaires to the greatest extent possible. In addition, fourth graders can be very literal in their interpretation of items and should not be expected to respond to an item’s underlying intent.

It was also discovered that students, in particular fourth graders, have a very hard time accurately estimating time and reporting on the frequency of behaviors they or their teacher engage in in class. Therefore, these items should be avoided to the greatest extent possible.

Finally, items presented in a list format can present problems. Respondents often forget (or do not attend to) the item stem of these types of questions, and answer them as stand-alone items. Reinstating the context in the list questions by repeating part of the stem can avoid this type of item problem.

This research project was sponsored by the Education Statistics Services Institute (ESSI), Washington, D.C., USA.

Copies of the U.S. Department of Education, National Center for Education Statistics, Working Papers that describe the research studies upon which this paper is based are accessible through the following URLs:

<http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200119>

<http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200206>

## References

- Anderson, R. and Pichert, J. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, 17, 1–12.
- Edwards, W., Levine, R., and Cohany, S. (1989). Procedures for validating reports of hours worked and for classifying discrepancies between questionnaire reports and validation totals. *Proceedings of the American Statistical Association*.
- Fisher, R. and Chandler, C. (1984). Dissociations between temporally-cued and theme-cued recall. *Bulletin of the Psychonomic Society*, 22, 395–397.
- Fisher, R. and Quigley, K. (1992). Applying Cognitive Theory in Public Health Investigation: Enhancing Food Recall with the Cognitive Interview. In J. Tanur (ed.), *Questions About Questions: Inquires into the Cognitive Bases of Surveys*. New York: Russell Sage Foundation, pp.154-169.
- Jobe, J. and Mingay, D. (1991). Cognitive and Survey Measurement: History and Overview. *Applied Cognitive Psychology* 5, 175–192.
- Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.
- Krosnick, J. (1999). Survey Research. *Annual Review of Psychology*, 50, 537–567.
- Levine, R., Huberman, M., Allen, J., and DuBois, P. (2001). *The Measurement of Home Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eighth Graders to Questionnaire Item and Parental Assessment of the Invasiveness of These Items*. U. S. Department of Education, National Center for Education Statistics, NCES 2001-19.
- Roediger, H. and Payne, D. (1982). Hypermnnesia: The role of repeated testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 66–72.
- Schwarz, N. and Sudman, S. (1996). *Answering Questions*. San Francisco: Jossey-Bass Publishers.
- Sudman, S., Bradburn, N., and Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.
- Tourangeau, R. (1984). Cognitive Sciences and Survey Methods, in T. Jabine, M. Straf, J. Tanur, and R. Tourangeau (eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*. Washington, DC: National Academy Press, pp. 73–100.
- Tulving, E. and Thomson, D. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352–373.
- Willis, G, Royston, P., and Bercini, D. (1991). The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires. *Applied Cognitive Psychology* 5, 251–267.
- Willis, G., Stinson, L., and Welniak, E. (1999). Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques, in M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur, and R. Tourangeau (eds.). *Cognition and Survey Research*. New York: John Wiley and Sons, Inc., pp. 133–153.