

Running head: COGNITIVE LABS AND TEST DEVELOPMENT

The Applicability of the Cognitive Laboratory Method to the
Development of Achievement Test Items

Christine Andrews Paulsen

Roger Levine

American Institutes for Research

April 23, 1999

Paper presented in *Research in the Development of Tests and Test Items*, C. Welch (Chair),
B. Zumbo (Discussant), at the annual meeting of the
American Educational Research Association, Montreal

Please send inquiries to Christine Paulsen, AIR, Center for Educational Assessment,
1000 Thomas Jefferson Street NW, Washington, DC 20007 –
V: 202-944-5443 -- F: 202-944-5454 – cpaulsen@air-dc.org

Abstract

This study explores the utility of the cognitive laboratory, and specifically the think aloud technique, for the development of achievement test items in 4th grade reading and 8th grade mathematics. While cognitive laboratory methods have been used over the past two decades to reduce the measurement error associated with surveys, this study represents the first systematic use of cognitive laboratories for test development.

Four hundred and thirty students participated in the cognitive laboratories, which were held at six urban and suburban sites across the country. The self-selected sample of students included 234 4th graders and 196 8th graders. Five hundred and eighty-four test items were evaluated in the laboratories, including multiple choice, short constructed response, extended constructed response, and drawn or gridded response items.

We found that a large proportion of the items were problematic; 80% of reading items and 74% of mathematics items exhibited problems. The most common problem types identified in the labs were problems with unclear language or contexts, and unclear or incomplete scoring rubrics. The cognitive labs, when used in combination with expert reviews, appear to provide significantly greater information on item problems than either method alone in mathematics. In both math and reading, the combination of cognitive labs and expert reviews resulted in greater identification of typographical and text editing problems. The cognitive labs were also especially useful at identifying language/contextual and rubric problems in mathematics items. These findings show that the cognitive laboratory method and the think aloud procedure appear to be useful tools for studying the processes and skills used by students when answering test questions.

Introduction

When students respond to items on achievement tests, one potential source of error is nonsampling, or measurement, error. Measurement error acts to reduce the reliability, and subsequent validity, of the test that it affects. Historically, test developers have mainly utilized psychometric tools in order to improve the validity and reliability of their achievement tests. However, such approaches assume that when a student answers a question correctly, the student has mastery of the concept being measured. Conversely, it is assumed that when a student answers a question incorrectly, it is because the student does not have mastery of the concept.

Such approaches cannot discern between those students who actually have mastery and those students who answered the item correctly because there was a “clue” hidden somewhere in the question, or for other non-mastery reasons. Similarly, statistical summaries of item performance cannot provide information about students who answer questions incorrectly because of lack of mastery versus students who answer questions incorrectly for the wrong reasons. For example, say a student is given an algebra item that requires them to compute the amount of an incentive a salesperson will receive for selling a certain number of boxes of cookies. It is possible that some students will be unfamiliar with the concept of an “incentive.” In those cases, students may answer the question incorrectly – not because they lack mastery of algebra, but because they are confused about the question itself.

When achievement tests are developed, items are routinely subjected to review in pilot and field tests, small scale try-outs, and panels of content area experts or bias reviewers. As discussed, typical psychometric approaches to item development provide essential information on factors such as item difficulty, but they are incapable of providing insight into the reasons behind how and why students answer questions differently. Similarly, expert reviews provide

essential information on issues such as potential bias problems or content coverage, but because they read each question as an expert, rather than as a novice test-taker, expert reviews lack important information critical to the test development process. Experts can miss “clues” or alternate valid response choices simply because they know what the right answer should be.

While both psychometric approaches and expert reviews are essential parts of the item/test development process, they provide only part of the whole story. There is a clear need for an item development procedure that can capture information about the reasons behind student performance on achievement test items, thereby enabling test developers to determine whether students are using the knowledge and skills intended for a given item. Such information will enable further refinement of test items, and will ultimately enhance the reliability and validity of achievement tests.

Based on work in the field of survey research, the cognitive laboratory method and think aloud procedure are new tools currently being explored for informing test development. For decades, researchers in the field of survey methodology have been developing techniques aimed at reducing measurement error. In the early 1980’s, collaborative research involving cognitive psychologists and survey methodologists yielded the cognitive lab approach for improving the reliability of items.

The cognitive laboratory method utilizes procedures intended to assist in understanding respondents’ thought processes as they respond to questions. This approach is used regularly by federal agencies involved in the collection of survey data. This includes the Bureau of the Census, Bureau of Labor Statistics, National Center for Health Statistics, and the National Center for Education Statistics, among others. Typically, the laboratories have been used to improve

surveys, rather than achievement tests. Therefore, there is currently no available research data that evaluates the utility of the cognitive lab approach for improving item development.

The objective of this study was to explore the applicability of the cognitive laboratory method, and in particular the use of the “think aloud” procedure, to the development of achievement test items in 4th grade reading and 8th grade mathematics.

Cognitive Labs and the Think Aloud Procedure

The evolution of cognitive labs and the think aloud procedure can be traced directly to paradigm shifts within the fields of psychology and survey methods. Early in this century, the field of psychology developed an increased focus on introspective methods for analyzing human behavior. Introspection was based on the notion that it was possible for trained psychologists to observe events that occurred in human consciousness – just as it was possible to observe events that took place in the outside world (van Someren, 1994). Verbal reports provided the main vehicle for data collection. However, these methods had weaknesses. First, introspective methodologists could not answer the question of whether the process of introspection itself was available to human consciousness. Second, experiments on human cognition could not be replicated because of the idiosyncratic nature of the observations. As the inherent weaknesses of introspective methods became clear, the majority of the scientific community turned away from introspection and cognitive research to focus more on behaviorism in the early 1930s (van Someren, 1994).

The shift to behaviorism led to the suspicion among some scientists that verbal reports were not “true” data. In their seminal work on the use of verbal reports as data, Ericsson and Simon (1980) argue that behaviorists have been “schizophrenic about the status of verbalizations as data (p. 216).” They argued that, on the one hand, the use of verbal responses was a very

common method in experiments. On the other hand, thinking aloud was often dismissed as a form of introspection or as a method that was only useful for generating hypotheses. Thus, during the rise of behaviorism, little attention was paid to development and refinement of the think aloud methodology.

Simultaneously, one of the greatest concerns in the emerging field of survey research was measurement error – including interviewer bias effects, the effect of question wording, and the effect of question order (Sudman, Bradburn, & Schwarz, 1996). The inaccurate polling that surrounded the 1948 election brought into sharp focus the possible effects of measurement error (it also revealed a need for better sampling techniques). However, without a theoretical framework for studying these potential sources of measurement error, the research on question wording and questionnaire design in the 1950s stalled.

During the 1960s, the evaluation of War on Poverty programs and other social experiments led to the increased use of surveys to study behavior, rather than opinions. Because funding was directly tied to the results of such evaluations, the accuracy of the data, and consequently sources of measurement error, again became critical concerns (Sudman, Bradburn, & Schwarz, 1996). Researchers began to turn their attention back to the issues of question wording and design.

During the 1970s, leaders in the field of survey research, Sudman and Bradburn, conducted research which conceptualized the “survey interview as a social system with two roles united by a common task (Sudman, Bradburn, & Schwarz, 1996, p. 9).” They proposed three sources of response effects: the interviewer, the respondent, and the task itself. Their research found that the largest effects were related to the task. Task variables such as question wording

were considered critical (Sudman, Bradburn, & Schwarz, 1996). Such research naturally led to a new focus on the cognitive processes participants use to respond to tasks.

These developments in the field of survey methods coincided with renewed interest in the field of psychology in studying internal cognitive processes. As computers became more commonplace, empirical studies of information processing using cognitive laboratories and the think aloud method became increasingly popular (van Someren, 1994). These studies also turned their focus toward more naturalistic settings, especially in the area of human memory.

Thus, the interest among survey methodologists in cognitive processes, and the desire among cognitive psychologists to apply their information processing models to more “everyday” environments set the stage for collaboration between the two fields. The first systematic effort to collaborate occurred in 1978 in Great Britain. The Royal Statistical Society and the Social Science Research Council organized this event in order to explore retrospective and recall data in social surveys (Jobe & Mingay, 1991).

In 1980, the Bureau of Social Science Research, with funding from the U.S. Bureau of the Census and the Bureau of Justice Research, held a conference to discuss the feasibility of using cognitive research to redesign the National Crime Victimization Survey (Sudman, Bradburn, & Schwarz, 1996). Two years later, the Committee on National Statistics (CNSTAT), of the National Academy of Sciences, formed a panel that produced a report and agenda for further collaboration between cognitive psychologists and survey researchers (Jobe & Mingay, 1991). In 1983, the same Committee sponsored a seminar “Cognitive Aspects of Survey Methodology,” in a further effort to solidify the interdisciplinary relationship between the two fields (Sirken & Hermann, 1996).

Simultaneously in West Germany, the Zentrum für Umfragen Methoden und Analysen (ZUMA) sponsored empirical studies of cognitive processes related to questionnaires (Jobe & Mingay, 1991). ZUMA's research and the CNSTAT effort, are both credited with providing the impetus for a major interdisciplinary effort between the fields of cognitive and survey research (Sudman, Bradburn, & Schwarz, 1996; Jobe & Mingay, 1991).

The interdisciplinary work has continued, and several U.S. Government-sponsored cognitive laboratories currently exist. These include the U.S. Bureau of the Census, the Bureau of Labor Statistics, and the National Center for Health Statistics (Wolfgang, Lewis, & Vacca, 1994; Moore, Marquis, & Bogen, 1996; Johnson, O-Rourke, Chavez, Sudman, Warnecke, Lacey, & Horm, 1996; Wellens, 1994; Nolin, & Chandler, 1996; Herrmann, 1994). In addition, several universities and survey research centers in the United States and Europe have cognitive laboratories.

Theoretical Framework

Current theoretical models for understanding how participants respond to questions evolved mainly from the work of Sudman, Bradburn, and Schwarz (1996). Their model was developed in response to the need for improving the accuracy of information collected by survey questions. The model provides a framework for understanding how participants interpret and respond to questions, and enables researchers to understand whether participants actually answer questions as the survey developers intend. For the purpose of studying the use of the cognitive lab method for test development, we adapted Sudman, Bradburn, and Schwarz's survey response model to take into consideration the cognitive processes and skills students use when responding to math and reading items. Our model provides a framework for understanding how students

respond to test items, and enables us to establish whether students rely on the cognitive processes and skills that were intended by the item writers.

According to the adapted response model, the first task encountered by students is the interpretation of the question in order to understand its meaning (both the literal and the pragmatic meaning). Second, students must be able to identify which activities the item writer intended for them to undertake. After participants determine what a particular item requires of them, they will need to recall relevant information from memory or remember how to apply a specific skill in order to produce an answer. Next, as part of the process of generating a response, participants must determine how to answer the question. In the case of multiple choice items, participants will be limited to the choices that exist. In the case of short essay items, participants must be able to construct an appropriate answer. How much detail the students think they should provide will either be determined by the test directions to the student, or by their own past experiences with similar items. After a participant has chosen an answer, they may want to change the answer or modify it based the expectations or pressures they perceive as a function of answering questions in an interview setting.

The cognitive lab provides information that allows us to determine why students answer questions in certain ways. In other words, we are testing the assumption that students answer questions incorrectly because they lack mastery of a subject area or skill. We are also testing the assumption that students answer questions correctly because they understand an item, and have mastery of the subject area or skill. Our model predicts that some items will trigger a “breakdown” in the response process. Some students will answer questions correctly or incorrectly for the wrong reasons. For example, a student may not know a word or phrase that is relevant to the test item, thereby inhibiting the student’s ability to comprehend the question.

Without cognitive lab data on such cases, test developers would be unaware that certain words or phrases were placing some students at an unfair disadvantage, or that some test items were unnecessarily complicated.

Methods and Procedures

Participants and Sites. Five hundred and eighty-four items, including 276 reading and 308 math items, were evaluated in the cognitive labs. These items were multiple choice (MC), short constructed response (SCR), extended constructed response (ECR), and gridded response items developed for a proposed national, standards-based assessment. Four hundred and thirty students participated in the cognitive labs, including 234 4th graders and 196 8th graders. The labs were conducted in laboratories at six urban and suburban sites across the country. The sites were chosen to maximize the probability that the student sample would represent national proportions of ethnic and economic populations. In addition to the cognitive labs, expert reviews of the proposed test items were conducted. Seven math and 7 reading experts participated in several review meetings conducted over a three month period. The experts reviewed items and made recommendations about whether to revise, drop or keep items “as is.”

Cognitive Lab Procedures and Instruments. Item-based interview protocols contained a set of item-specific prompts and probes designed to assist interviewers in understanding students’ answers to test questions (Ericsson & Simon, 1993). Prompts were general statements used to encourage students or to reinforce their use of the think-aloud procedure (e.g., *“You’re doing a great job thinking aloud.”*) General prompts were used until a response was stated, and then specific probes were used as necessary, to insure that the interviewer knew how the student arrived at his or her answer and to determine if the student had mastery of the construct(s) the item was purporting to measure.

The cornerstone of the cognitive lab interviews was the think aloud procedure, based on the work of Ericsson and Simon (1980). Before each interview began, students were trained in the procedure. The think aloud involved the following steps. First, a question or item was presented to a student. The student was asked to read the item, and given time to attempt to respond to it. As the student attempted an answer, they were instructed to verbalize their thoughts. Depending on the type of information that was communicated by the student, the interviewer used different probes or prompts to encourage students to provide more/different information. The goal of the think aloud activity was to obtain as much evidence as possible of the reasons behind student responses to test items.

Interviews lasted between one and a half to two hours, during which students completed between eight and 14 items. Interview information was recorded using videotape and a standard summary form. These forms recorded students' responses to each item, and described (with supporting evidence) any potential item problems that were discovered in the lab.

Research Questions and Analysis

The first research question was: What information can the cognitive labs provide about test items? To study this, we computed the proportion of item problems discovered in the math and reading items. Within each content area, we computed the proportion of item problems for each of the different item formats (i.e., multiple choice, short constructed response, extended constructed response, and gridded math items).

The second research question was: How does the information on potential item problems collected in the labs compare with the information on item problems collected by expert reviewers? We selected a random sample of 360 items, stratified by item format (e.g., multiple choice, short constructed response, etc.) to study the question of whether the cognitive labs provided useful information beyond that provided by expert reviewers. We content analyzed the

item problems found in three different types of review: cognitive lab reviews (based on analysts' syntheses of interview data), recommendations made by content experts for items evaluated in the labs (in other words, the content experts had access to lab data when they made their recommendations), and recommendations made by experts for items that did not go through the labs. The sample of items was split evenly between math and reading (the sample size was 180 for both content areas, containing 60 items from each review group).

Based on an analysis of the content of lab and reviewer data, we developed a typology of problems found. Item problems identified by the two groups included:

- typographical errors and problems with text that needed to be edited to make items more grammatically correct or easier to read;
- formatting problems, which usually meant the item needed to be re-organized so that different parts of the question were clearly highlighted, that the item had too many parts, or that the question format (MC, SCR, or ECR) was inappropriate;
- some items were considered too ambiguous, either in terms of the instructions provided to students or the rubrics were unclear about what was expected of students in order to receive full credit, in addition, some items were missing information that was essential to solving the problem or answering the question;
- incomplete or incorrect artwork or graphics;
- potential language problems – words or phrases that could confuse or distract students, extraneous content or confusing contexts that could mislead or distract students from the task;
- items in which an incorrect answer was specified, or an alternative correct answer is possible from among the distractors;
- items in which students could make a correct or full-credit response without having mastery of the concepts (“back-door”);
- items that were too difficult; and
- items that were too easy or considered “trivial.”

Results

Out of the 276 reading items that were evaluated in the cognitive labs, 222 (80%) exhibited at least one item problem. Prevalence of item problems differed for multiple choice (MC), short constructed response (SCR) or extended constructed response (ECR) items. Table 1 shows that while we found that 62% of the MC items exhibited at least one item problem for at

least one student, we found that 87% of the SCR items demonstrated at least one item problem for at least one student. Moreover, almost every ECR item (93%) exhibited a problem.

TABLE 1
Frequency of Reading Items with Problems by Item Format
All Items (n=276)

Item Format	Frequency of Item Problems			
	No Problems	1 Type of Problem	2-4 Types of Problems	Total
MC	32 (38%)	32 (38%)	21 (25%)	85 (100%)
SCR	18 (13%)	55 (38%)	72 (50%)	145 (100%)
ECR	3 (7%)	17 (37%)	26 (57%)	46 (100%)
Total	54 (20%)	103 (37%)	119 (43%)	276 (100%)

Most of the problematic items exhibited more than one type of problem. Of the 222 items that were identified as problematic, 54% exhibited two or more problems. In addition to exhibiting more total item problems overall, the SCR and ECR items were also more likely to display a greater number of types of problems than the MC items. Table 1 shows that 50% of the SCR items and 57% of the ECR items displayed at least two different types of problems, while only 25% of the MC items had two or more types of problems.

Of the 308 items reviewed in mathematics, item problems were discovered in 229 (74%) of the items. As with reading items, different item formats yielded differences in the extent of problems with math items. In fact, the breakdown of math item problems mimics that of reading. Table 2 illustrates the frequency of problem types. ECR items were more problematic than other item formats: 96% of all ECR items exhibited at least one problem type, compared with 79% of SCR items, 59% of multiple choice items, and 58% of gridded items.

As with reading items, the math ECR items were also more likely than other item formats to exhibit more than one problem. As Table 2 illustrates, 54% of all ECR items had at least two

different types of problems, while 41% of SCR items, 30% of multiple choice items, and 20% of gridded items exhibited at least two different types of problems.

TABLE 2
Frequency of Mathematics Item Problems by Item Format
All Items (n=308)

Item Format	Frequency of Item Problems			
	No Problems	1 Type Of Problem	2-4Types Of Problem	Total
Multiple Choice	26 (41%)	18 (13%)	19 (30%)	63 (100%)
SCR	31 (21%)	56 (38%)	60 (41%)	147 (100%)
ECR	2 (4%)	20 (42%)	26 (54%)	48 (100%)
Gridded	21 (42%)	19 (38%)	10 (20%)	50 (100%)
Total	79 (26%)	113 (37%)	116 (38%)	308 (100%)

Next, we explored the types of item problems identified in the labs, and how these compared to the types of problems identified in expert reviews. Tables 3 and 4 show the nine different types of item problems identified by our content analysis. These nine types of item problems were identified in both cognitive labs and expert reviews.

Tables 3 and 4 show that the most common problem discovered in the cognitive labs was items that contain extraneous language, and contrived or confusing contexts or language. Nineteen of the 64 (33%) item problems identified in the labs for math items, and 23 of the 49 (47%) problems identified in reading were of this type. The second most common problem type identified in the labs for both math and reading was the existence of unclear instructions and rubrics, and lack of correct answers (30% and 20%, respectively). The most common item problem discovered by expert reviewers was the need for text edits, and typographical errors (35% of math items and 28% of reading items reviewed by experts exhibited these problems). The expert reviewers also discovered a large proportion of items with language and rubric problems.

TABLE 3:
Comparison of Reading Item Problems Identified by Three Review Groups
(n = 180 items)

Type of Item Problem Identified	Frequency of Item Problem Identified ¹				χ^2 goodness-of-fit
	Expert Only (n=60)	Cog Lab Only (n=60)	Cog Lab & Expert (n=60)	Reading Totals (n=180)	
Text edits or typographical errors	12.5 (44%)	2 (7%)	14 (49%)	28.5 (100%)	8.98*
Formatting problems	2.5 (45%)	0 (0%)	3 (55%)	5.5 (100%)	2.82
Unclear instructions or rubrics, or no clear answer	8.3 (27%)	10 (32%)	13 (42%)	31.3 (100%)	1.09
Artwork or graphics problem	0 (0%)	0 (0%)	0 (0%)	0 (100%)	0.00
Extraneous information, contrived or confusing context or language	10.8 (20%)	23 (44%)	19 (36%)	52.8 (100%)	4.40
Incorrect answer specified, possible alternative correct answer exists	3.3 (19%)	9 (49%)	6 (33%)	18.3 (100%)	2.64
Possible back-door to correct answer	0.8 (17%)	1 (21%)	3 (62%)	4.8 (100%)	1.81
Too difficult	1.7 (36%)	2 (43%)	1 (21%)	4.7 (100%)	0.33
Too easy/trivial	5.0 (55%)	2 (22%)	2 (22%)	9.0 (100%)	1.98
Total Frequency of Problems	44.8 (29%)	49 (32%)	61 (39%)	154.8 (100%)	2.73

Note. Because an item can exhibit more than one item problem, the frequencies may add up to more than the actual number of items.

¹Cognitive lab items were actually reviewed by internal experts before being sent to the labs to ensure the items were “usable” in the lab setting. This review was not as intensive as the expert review, but did result in some item problems being identified before the items reached the labs. This internal review did not occur before external experts reviewed items. Therefore, the frequency of item problems reported by experts is inflated. Internal item reviewers estimate they identified and corrected about 17% of the reading item problems before the items were evaluated in the labs. We therefore assume that the frequency of item problems identified by external review is inflated by approximately 17%. The Expert Only column has been adjusted for this inflation (by a factor of .83).

* $p < .05$.

As shown in Table 3, the cognitive labs used in combination with expert reviews, appear to identify a significantly greater proportion of items in need of text edits and typographical changes ($\chi^2_{(df=2)} = 8.98, p < .05$).

Table 4 shows that overall, the cognitive labs and expert reviews combined provide more information on mathematics item problems than either method alone ($\chi^2_{(df=2)} = 20.60, p < .05$).

Moreover, the combination of labs and expert reviews led to the discovery of a greater number of items requiring text edits or typographical changes ($\chi^2_{(df=2)} = 20.95, p < .05$). The cognitive labs

also led to the discovery of a greater number of problems with unclear instructions or rubrics

($\chi^2_{(df=2)} = 8.70, p < .05$), and language or contextual problems ($\chi^2_{(df=2)} = 9.38, p < .05$).

TABLE 4:
Comparison of Math Item Problems Identified by Three Review Groups
(n = 180 items)

Type of Item Problem Identified	Frequency of Item Problem Identified ¹				
	Expert Only (n=60)	Cog Lab Only (n=60)	Cog Lab & Expert (n=60)	Math Totals (n=180)	χ^2 goodness-of-fit
Text edits or typographical errors	12.3 (26%)	5 (11%)	30 (63%)	47.3 (100%)	20.95*
Formatting problems	2.6 (15%)	8 (45%)	7 (40%)	17.6 (100%)	2.76
Unclear instructions or rubrics, or no clear answer	5.3 (12%)	19 (44%)	19 (44%)	43.3 (100%)	8.70*
Artwork or graphics problem	6.2 (34%)	5 (28%)	7 (39%)	18.2 (100%)	0.33
Extraneous information, contrived or confusing context or language	5.7 (12%)	21 (45%)	20 (43%)	46.7 (100%)	9.38*
Incorrect answer specified, possible alternative correct answer exists	0.9 (31%)	0 (0%)	2 (69%)	2.9 (100%)	2.09
Possible back-door to correct answer	0.9 (23%)	3 (77%)	0 (0%)	3.9 (100%)	3.68
Too difficult	1.8 (37%)	2 (42%)	1 (21%)	4.8 (100%)	0.34
Too easy/trivial	0 (0%)	1 (100%)	0 (0%)	1 (100%)	2.00
Total Frequency of Problems	35.6 (19%)	64 (35%)	86 (46%)	185.6 (100%)	20.60*

Note. Because an item can exhibit more than one item problem, the frequencies may add up to more than the actual number of items.

¹Cognitive lab items were actually reviewed by internal experts before being sent to the labs to ensure the items were “usable” in the lab setting. This review was not as intensive as the expert review, but did result in some item problems being identified before the items reached the labs. This internal review did not occur before external experts reviewed items. Therefore, the frequency of item problems reported by experts is inflated. Internal item reviewers estimate they identified and corrected about 56% of the math item problems before the items were evaluated in the labs. We therefore assume that the frequency of item problems identified by external review is inflated by approximately 56%. The Expert Only column has been adjusted for this inflation (by a factor of .44).

* $p < .05$.

Discussion

The large proportion of problematic items is not surprising, given that the items used in the labs were often in very early stages of development. In fact, these findings illustrate how sensitive the labs can be in detecting problems in items that have not been thoroughly reviewed and revised. The labs reveal not only the extent of item problems, but also which item formats are more problematic (e.g., SCRs and ECRs). Such information can be useful to test developers as they develop training materials for item writers, or plan for item attrition among different item types.

The cognitive labs appear to provide important supplemental information to the expert reviews. When used in combination with expert reviews, the labs can validate and expand findings about potential item problems. This is especially true for typographical or text editing problems in both math and reading. The labs also appear to provide important additional information beyond expert reviews on language/contextual issues and rubrics in the subject of mathematics. While expert reviewers play an essential role in determining whether specific items cover particular content areas well or whether items are appropriate for students they are adult experts and may not be able to anticipate what could be confusing to children. The strength of cognitive labs is their ability to show how particular items perform with actual students. Results from the labs provide test developers with an understanding of how students approach test items and enable developers to determine whether students are using the strategies and processes expected by item writers.

Cognitive labs appear to represent a viable alternative for test developers who need information on test item quality. Whether the labs are worth the extra cost is an issue that is beyond the scope of this paper, but is an important one for test developers to consider. What the

labs can offer, that the expert reviews cannot, is an opportunity to see how items will perform with actual students.

These findings also demonstrate that children, even as young as 4th graders, are fully capable of participating in cognitive labs. Labs that employed 4th grade students identified just as many problems with reading items as did the experts reviewing reading items.

This study is important because it represents the first systematic attempt to explore the use of cognitive labs, and the think aloud procedure, for the test development process. It is critical to understand whether the cognitive lab methodology can be applied to the area of test development in an effort to improve the reliability, and subsequent validity, of tests – especially high stakes tests. Collecting feedback from individual students as they attempt to answer test items adds an extra level of validity to the test, and helps to ensure that test items are fair. In other words, the cognitive lab enables test developers to understand whether each item is measuring what they intended to measure.

This study is also important because it provides concrete information about the effectiveness of the procedures and techniques used in the labs. Previous cognitive lab study designs have been based largely in theory, as well as the experiences of survey researchers. These studies were unable to benefit from systematic cognitive lab studies that were conducted in the field of test development, because none existed. This current study fills that gap, and provides cognitive lab researchers with real evidence about procedures and techniques and how these tools can inform the test development process.

References

- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. Psychological Review, 87(3), 215-251.
- Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data. Cambridge: MIT Press.
- Herrmann, D. (1994, December). The contributions of the NCHS collaborative research program to memory research (Cognitive Methods Staff Working Paper Series, No. 14). Hyattsville, MD: Office of Research and Methodology, National Center for Health Statistics.
- Jobe, J. B. & Mingay, D. J. (1991). Cognition and survey measurement: History and overview. Applied Cognitive Psychology, 5, 175-192.
- Johnson, T. P., O-Rourke, D., Chavez, N., Sudman, S., Warnecke, R. B., Lacey, L., & Horm, J. (1996). Cultural variations in the interpretation of health survey questions. In R. Warnecke (Ed.). Health survey research methods: Conference proceedings. Hyattsville, MD: National Center for Health Statistics, 57-62. (DHHS publication number PHS 96-1013)
- Moore, J. C., Marquis, K. H., & Bogen, K. (1996). The SIPP cognitive research evaluation experiment: Basic results and documentation. Washington, DC: U.S. Bureau of the Census.
- Nolin, M., & Chandler, K. (1996). Use of cognitive laboratories and recorded interviews in the National Household Education Survey. (Report No. NCES-96-332). Rockville, MD: Westat. (ERIC Document Reproduction Service No. ED 401 337)
- Sirken, M., & Herrmann, D. (1996). Relationships between cognitive psychology and survey research. Proceedings of the American Statistical Association, Section on Survey Research Methods, 245-247.

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). Thinking about answers: The application of cognitive processes to survey methodology. San Francisco: Jossey-Bass.

van Someren, M. W. (1994). The think aloud method: A practical guide to modelling cognitive processes. San Diego, CA: Academic Press.

Wellens, T. (1994). The cognitive evaluation of the nativity questions for the Current Population Survey. Proceedings of the American Statistical Association, Section on Survey Research Methods, 1204-1209.

Wolfgang, G. S., Lewis, P. J., & Vacca, E. A. (1994). Cognitive research for the 1997 census agricultural report form. Proceedings of the American Statistical Association, Section on Survey Research Methods, 503-508.