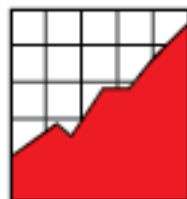**Technical Report 44**

Using the Think Aloud Method (Cognitive Labs) To Evaluate Test Design for Students with Disabilities and English Language Learners

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

In collaboration with:

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)

Technical Report 44

Using the Think Aloud Method (Cognitive Labs) To Evaluate Test Design for Students with Disabilities and English Language Learners
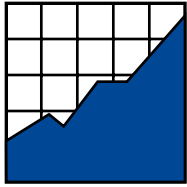
Christopher J. Johnstone • Nicole A. Bottsford-Miller •
Sandra J. Thompson*

August 2006

* Dr. Thompson was a Research Associate at NCEO when working on this report. She passed away December 2005 after a career of improving outcomes for students with disabilities.

NCEO Core Staff

Deb A. Albus	Michael L. Moore
Manuel T. Barrera	Rachel F. Quenemoen
Christopher J. Johnstone	Dorene L. Scott
Jane L. Krentz	Karen E. Stout
Kristi K. Liu	Martha L. Thurlow, Director
Ross E. Moen	

Executive Summary

Finding ways to improve the design of large-scale tests is a timely issue. Recent changes in Federal legislation (including the No Child Left Behind Act of 2001) have placed greater emphasis on accountability via large-scale tests. Students who have previously been exempted from large-scale assessments, including students with disabilities and English language learners, are now expected to reach achievement levels comparable to their non-disabled or English proficient peers. Schools are held accountable for their performance, and their scores are reported publicly. With such high stakes placed on large-scale assessment, there is a critical need for states to have valid information about how the design of assessments affects student test performance.

This report provides information on the use of "think aloud methods" to detect design issues in large-scale assessments. Various design problems may introduce construct-irrelevant variance or hinder students from showing what they know on assessments. Our research included a variety of students, including students with learning disabilities, students with hearing impairments, students with cognitive disabilities, English language learners, and students without disabilities who were proficient in English. In this project, we asked students to "think out loud" when solving mathematics large-scale assessment items. The sentences that students uttered produced data that led us to believe that think aloud methods, under certain circumstances, can successfully detect design issues. Specifically, we found issues related to unclearly defined constructs, inaccessibility of items, unclear instructions, incomprehensible language, and illegible text and graphics. To this end, think aloud methods appear to be a useful strategy in the design and refinement of large-scale assessments.

Think aloud methods, as we designed them, were not effective for students with cognitive disabilities. This population had great difficulty in producing the language needed to explain problem-solving processes and may require additional research accommodations to help them participate in think aloud research. All other groups sufficiently participated in research activities. Think aloud methods also did not produce informative data for very difficult mathematics items because students had difficulty verbalizing their thoughts while solving problems. Despite shortcomings found in this study, the think aloud method appears to be an effective way to determine the effects of item design for a wide variety of students (with the exception of students with cognitive disabilities) and for items with low to moderate difficulty levels.

Table of Contents

Introduction

Recent research on test design is currently being conducted from a universal design framework (Dolan, Hall, Banerjee, Chun, & Strangman, 2005; Ketterlin-Geller, 2005), which states that test design should be accessible and understandable to a wide variety of users (including students with disabilities and English language learners). According to Thompson, Johnstone, and Thurlow (2002), elements of universally designed assessments include (1) inclusive test population; (2) precisely defined constructs; (3) accessible, non-biased items; (4) amenable to accommodations; (5) simple, clear, and intuitive instructions and procedures; (6) maximum readability and comprehensibility; and (7) maximum legibility. Research indicates that test designers can create assessments that are more accessible to diverse students by designing items using elements of universal design (Johnstone, 2003). They can also minimize construct-irrelevant variance (Haladyna, Downing, & Rodriguez, 2002) by adhering to effective design strategies. Such design features may increase the validity of information that can be gleaned from test data.

Research from the 1980s (Grise, Beattie & Algozzine, 1982) to the present (Dolan et al., 2005; Johnstone, 2003) has attempted to clarify design issues by demonstrating how specific design improvements in tests can affect student performance. Such research is limited, however, because it only provides information on the final product of student responses on tests. Little is currently known about how the design of a test relates to student processing of items (i.e., what happens to create a particular outcome). To understand and make predictions about test design effects on student processes involved in test taking, research design must be set up to tap the cognitive processes of students while they work through test items. Researchers can access cognitive processes in a concrete fashion by requesting students to verbalize as they think, or "think aloud."

The purpose of this report is to focus on the Think Aloud Method (Cognitive Laboratory) research methodology to detect design issues in large-scale tests, based on a framework of universal design. To this end, we used Thompson et al.'s (2002) *Elements of Universally Designed Assessments* (Table 1) as a template for detecting possible design issues. We describe the methodology in general and evaluate its effectiveness for finding design issues in tests for students with disabilities, English language learners, and English proficient students without disabilities. Finally, we discuss limitations and future directions for this methodology, particularly for students with disabilities with whom this methodology has not been used extensively before.

The Think Aloud Method: Background

The use of verbalizations as indicators of cognition is a decades-old data collection technique. Psychologist Karl Duncker (1945) originally described think aloud verbalizations as "produc-

Table 1. Elements of Universally Designed Assessments

1.	Inclusive assessment population
2.	Precisely defined constructs
3.	Accessible, non-biased items
4.	Amendable to accommodations
5.	Simple, clear, and intuitive instructions and procedures
6.	Maximum readability and comprehensibility
7.	Maximum legibility

tive thinking" and a way to understand his subjects' development of thought. Fifty years later, Ericsson and Simon (1993), authors of the book *Protocol Analysis: Verbal Reports as Data*, posited that think aloud data collection is a valid method for researching cognitive processes. According to the authors, think aloud methods draw on thoughts in the short-term memory of subjects. Because all cognitive processes travel through short-term memory, the conscious thoughts of the subject can be reported at the time they are processed. According to Ericsson and Simon, the cognitive processes that generate verbalizations ("think alouds") are a subset of the cognitive processes that generate behavior or action.

There are both advantages and disadvantages to using information drawn from think aloud data. Collecting data from the short-term memory is preferable because thoughts generated from the long-term memory of subjects are often tainted by perception. Ericsson and Simon (1993) stated that once information enters the long-term memory, subjects may incorrectly describe the processes they actually used. Verbalizations that take place concurrently with cognitive processes are largely independent of interpretation on the part of the subject (Van Someren, Barnard, & Sandberg, 1994).

Conversely, gathering data in real-time can be problematic because think aloud utterances are often incoherent (Ericsson & Simon, 1993). More articulate responses can generally be drawn from interviews, which take place after the think aloud protocol is completed. Branch (2000) identified disadvantages of the think aloud method. She found that the cognitive load of problem solving and speaking may be too difficult for some subjects. This problem can be mitigated by using retrospective data. Branch (2000) and Fonteyn, Kuipers, and Grobe (1993) all found that asking post-process questions to subjects provided valuable information that made think aloud data easier to understand and interpret.

A two-step process appears to be a practical approach to think aloud techniques. In this method, researchers first collect data in real time, asking subjects to think aloud. During the first step, researchers probe subjects as infrequently as possible because subjects are easily distracted dur-

ing problem-solving activities (Ericsson & Simon, 1993). When silences continue for several seconds, researchers merely probe the subject to "keep talking." Neutral cues such as "keep talking" encourage subjects to think aloud but do not bias the data by adding external ideas to the internal processes of subjects.

Once the think aloud process is complete, the second step of this method is to ask follow-up questions. Answers to these questions are not the primary data source, but can supplement any unclear data derived from think aloud techniques. Such questions may also be useful for subjects who are unable to meet the cognitive demands of thinking aloud while problem solving (Branch, 2000).

Using Think Aloud Methods for Evaluating Test Design

Think aloud protocols are becoming more common in educational research due to the richness of data that potentially can be derived from the methodology. Pressley and Afflerbach (1995) used think aloud protocols extensively in their research on how readers engaged in a variety of literary activities. The protocols the readers produced in response to the method provided the researchers with valuable data on how readers of varying abilities adjust to different types of text.

Kopriva's (2001) work with English language learners examined assessments using think aloud methods. She recommended that all test designers use think aloud methods to better understand test design and its effects on student test-taking processes. According to Kopriva, verbalizations used for think aloud data provide valuable insights into the following:

- Student understanding of constructs
- Student skill level
- Relevance of items to student life experience (see also Rothman, 1993)
- Relevance of items to content taught

Such information relates to research on universal design of assessments (Thompson et al, 2002) that encourages test-makers to be mindful of:

- Construct fidelity
- Potential bias
- Possibilities for accommodation
- Comprehensibility of instructions
- Comprehensibility, readability, and legibility of items

Using think aloud data to examine test-taking processes, with careful scrutiny on design features,

can aid test producers in understanding how design affects student performance.

Think aloud methods, within the context of assessment, also have challenges. Leighton (2004) noted that think aloud methods can be effective for understanding item qualities for assessments, but recognized several limitations to the method, such as the decreased ability of researchers to gain meaningful data from items that are too simple or too challenging for students. Pressley and Afflerbach (1995) described the potential and challenges for think aloud methods, noting that the richness of language (or lack thereof) "are the greatest assets and liabilities of the verbal reporting methodology" (p. 2).

Research in related fields, however, has demonstrated that using think aloud data can lead to better designed products. Shriver (1984, 1991), for example, used think aloud data to improve readability of written documents. Likewise, Camburn et al. (2000) and Nolan and Chandler (1996) conducted think aloud experiments during the pilot stages of survey development and used data to improve the readability and accessibility of surveys. Think aloud methods are by no means meant to replace other assessment evaluation techniques (sensitivity reviews, statistical analysis of results, etc.) but may provide otherwise untapped information about test design and student performance.

Sampling When Using the Think Aloud Method

During think aloud studies, research subjects must spend several minutes (sometimes more than one hour) working their way through protocols. Because of the labor-intensive nature of this method, the sample size involved in the research is necessarily small. Small numbers, however, do not indicate small data sets. The research process is intensive, so small sample sizes still can provide valid information. Nielson (1994), for example, suggested that sample sizes as small as five participants will yield sufficient information about problem solving behavior.

Furthermore, unlike large questionnaire or psychometric research projects, samples are not chosen at random for think aloud protocols. Think aloud sampling is purposeful. Subjects are chosen as representatives of particular subsets of students deemed important to the project. Often these subsets have low incidence rates in the general population. Secondary groups that more closely align with national demographics are often also selected for think aloud studies for comparison purposes (Kopriva, 2001). The research project presented in this paper focused on students with disabilities and English language learners. Students without disabilities who were English language proficient were chosen as the comparison group.

To ensure valid information during our think aloud study, we sought to obtain a 5−10 student sample per group. This number far exceeds the overall sample size deemed appropriate by Nielson (1994), but would give us meaningful information within groups and between groups.

Students were chosen to represent the following groups:

- Students with learning disabilities (LD)
- English language learners (ELL)
- Students who are deaf or hard of hearing (Deaf/HH)
- Students with cognitive disabilities (CogD)
- Students without disabilities who were proficient in English (comparison group)

Table 2 shows the number of subjects per cell for the primary analysis. The only cell that did not have five subjects was eighth-grade students with cognitive disabilities who took statewide assessments (the population size of this subgroup is very low in the school district where research took place).

Table 2. Sample Size by Sub-group

Grade	Learning Disabled (n)	Deaf/Hard of Hearing (n)	Cognitive Disability (n)	English Language Learners (n)	Non-Disabled (n)
4	9	10	6***	7	10
8	10*	11**	3	9	10

*Includes one 9th grader
**Includes four 7th graders
***Includes one 5th grader

Data Collection

Data were collected in elementary and middle schools on the outskirts of a large, urban area in the U.S. midwest. Protocols used were selected prior to any fieldwork. Because the analysis focused on large-scale assessment items, statewide test data were used to determine what large-scale assessment items students had particular difficulty solving. Using statistical techniques (item total correlation, item ranking, and pass cross-tabulation), researchers determined a priori which items were particularly problematic for target subjects (students with disabilities and English language learners). With permission from the state and test publisher, six items were reproduced and used as protocols for student think aloud techniques. Reproduction of items ensured that each student had a working copy during think aloud activities.

Field work began when we met with students individually. Each student was asked to sit down at a table. The researcher then explained think aloud procedures and demonstrated the process of verbalizing while thinking. The actual script that we used to explain the process is found in Figure 1.

Figure 1. Think Aloud Protocol Script

"We are interested in how students solve problems on tests, so we want to ask you and other students to solve some test problems for us and let us listen to how you do that. We are not as interested in the answer you come up with as we are with how you are thinking about the tasks."

Notice the phrasing is general and honest about our interests and respectful of the contribution each student can make to tests for students across the country. Students should not feel the slightest sense of being judged or of having to obtain any particular types of results. Once they do, it affects their behavior and introduces a bias.

Ask the student to "parrot" back what he or she was told about today's session by the recruiting person or teacher. Often, you will find that the student has been given information that is biasing and can affect the session. You need to find it in order to rectify it.

"What were you told we were going to do today?"

Be curious about what students do and why. Also tell the student that you will be videotaping the session and let him/her know when you turn on the camera.

"What you say is really important, so we are going to run this camera to make sure that we don't forget anything."

Provide practice

Give each student a practice task to familiarize him or her with thinking aloud while working through a task. First you solve a problem and then ask the student to solve one. (The camera is not turned on for the practice.) Give the following instruction:

"I'm going to think out loud while I solve this problem. That means I'm going to say everything that goes through my mind." (Complete problem while thinking out loud.)

"Now I'm going to ask you to solve a problem the same way. Just say everything that goes through your mind while you solve the problem."

"I am not as interested in the answer to the problem as much as how you are thinking about the task. Do you have any questions about what we just did?"

After we explained instructions and provided a short demonstration of how to verbalize, students engaged in a sample exercise to practice verbalizing their thoughts. We used a hidden picture activity to model and practice thinking aloud with students. Students practiced until they understood and could think aloud clearly and then were given instructions on the research protocol.

After completing the instructions, we asked students whether they had any questions, then watched students as they worked through problems. We cued students only when they were silent for 10 consecutive seconds. If students verbalized infrequently, we reminded them to "keep thinking aloud" or "keep talking." While students were thinking aloud, we remained silent to

avoid disrupting their thought patterns (Ericsson & Simon, 1993). When students completed an item, we asked follow-up questions for clarification.

There was no script for follow-up questions. Rather, researchers asked questions based on events that arose during the think aloud protocol. For example, we may have asked a process question ("How did you solve that?") when the student did not adequately verbalize. Or we may have asked a design question ("Was there anything that confused you?") when a student spent several minutes on a sub-section of an item. Table 3 demonstrates the introspective and retrospective data collected during these protocols.

Table 3. Types of Data

Type of Data	Example
Process	Student written material that demonstrates problem-solving process
Introspective	Student thoughts as they attempted to solve items
Retrospective	Student perceptions of items after items were completed

During data collection one person silently observed while the other conducted the think aloud process. The observer noted overall themes and specific events of the protocol. In addition, we video and audio taped all protocols using a tripod-mounted video camera. The observer ensured that video and audio equipment was functioning properly during the sessions. Capturing the think aloud protocols on both video and audio tape allowed us to review verbal data at a later date in more detail and to view the nuances present (Fonteyn, Kuipers, & Grobe, 1993) during the test-taking process.

Analysis of Data

We used two methods of coding data for this project. First, a product analysis (Van Someren, Barnard, & Sandberg, 1994) was completed to determine which items were answered correctly and incorrectly. The second (more intensive) analysis was a process analysis of the verbalizations students made while attempting test items. This second analysis was completed by watching video tapes and coding the major processes of test item completion while reviewing protocols.

We began the coding process by noting which behaviors and process information were sought a priori (Ericsson & Simon, 1993). Such a determination was thought to facilitate finding variation between students. Decisions were made by labeling the target task first (in this case, test taking), then analyzing the task into sub-components (Nielson, Clemmensen, & Yssing, 2002).

Sub-components of the test taking process were drafted onto a coding form for use when analyz-

ing think aloud data captured on video (Van Someren et al., 1994). Several iterations of coding sheets were drafted before a final version was chosen. Coding sheets were examined several times to ensure that they met the needs of this project and maintained established norms for think aloud studies.

The coding sheet designed by the research team contained a checklist for noting various subject behaviors and thoughts throughout the test-taking process (i.e., coding sheets provided space for researchers to examine all steps of the test-taking process independently, from reading the item to solving the problem). The final coding sheet also contained spaces for direct quotes, assertions, and actions that took place during the think aloud process. It is shown in Figure 2.

Coded data were analyzed both quantitatively and qualitatively. First, overall responses were coded for various design features. Frequency counts were tallied for correct/incorrect responses, distraction, problem-solving related issues, and reading-related issues. Sub-groups of the total sample (cells) were analyzed to determine whether differences existed in each of the categories (prompts/assistance, directions at top of page, item reading, etc.). From all these data, descriptive statistical summaries were produced (Kopriva, 2001).

Qualitative information was analyzed using qualitative research methods (Bogdan & Biklen, 1992), that is, all direct quotes from students were transcribed and coded into generalized themes that represented actual events (Van Someren et al., 1994). To ensure that information derived from coding was reliable, numerous raters were used during the data coding phase (including raters who collected data, raters who knew about data collection but did not participate, and raters who were new to the project at the analysis phase). All tapes were reviewed by one of two main raters. Every tenth tape was reviewed by four raters and checked for consensus.

Once coding was completed, an additional analysis was conducted using Thompson et al.'s (2002) *Elements of Universal Designed Assessments*. For each item, qualitative results were compared to *Elements* to determine whether think aloud methods could detect specific design issues.

Results

Because of test security issues, we are not able to report results for individual items. Rather, in this section we have organized the results of the items by universal design element. In total, students reviewed 12 items (6 fourth-grade items and 6 eighth-grade items). Results reported below describe both grade levels in order to demonstrate how think aloud methods highlighted design issues that sometimes varied by grade level.

The think aloud method appears to be best able to directly address Elements 2, 3, 5, 6, and 7 (accessible, non-biased items; simple, clear, and intuitive instructions and procedures; maxi-

Figure 2. Think Aloud Data Coding Sheet

Universally Designed Assessments – "Think Aloud" Study

Student ID # ___   Researcher Initials ___   Item # ___

Grade: 4 8   Describe Item: _____

Describe Researcher Introduction (if included on video)

Prompts/Assistance				
___	No Prompts	___ Researcher	___	Point To Item
___	Other	___ Teacher	___	Paraphrase Directions
		___ Interpreter	___	Paraphrase Item

Describe Interaction with Student

Directions at Top of Page			
___	NA	___ Student Read Aloud	___ Researcher Read Aloud
___	Other	___ Student Read Silently	___ Signed by Interpreter
		___ Student Skipped	___ Reader / Signer Skipped

Describe Reading / Skipping Directions

Item Reading			
___	Other	___ Student Read Aloud	___ Researcher Read Aloud
		___ Student Read Silently	___ Signed by Interpreter
		___ Student Skipped	

Describe Reading / Skipping

Reading Fluency		
___	NA (Student did not read item aloud)	___ Student read all words correctly
		___ Student mispronounced some words
		___ Student had difficulty with many words

List words mispronounced

Researcher asked follow-up questions   ___ Yes   ___ No

Describe follow-up questions and student responses

Problem Solving		
		___ Correct process for solving problem
___	Incorrect problem solving process	___ Appeared to guess
___	Not Apparent	___ Did not attempt to solve problem

Describe problem solving process

Student was distracted by something on the page   ___ Yes   ___ No   ___ Not Apparent

Describe distraction

*Add observer comment on back (O.C.) and note students to use as examples.

mum readability and comprehensibility; and maximum legibility). In this study, we were not able to assess Element 1 (inclusive testing population). Element 1 refers to larger patterns of student involvement in statewide assessment. Element 4 (amenable to accommodations) was not formally addressed in this study because we did not have braille forms from which to gauge if tests would be equivalent in different formats. We did, however, gather anecdotal evidence on the issues related to sign language interpretation. This information will be used for a more formal study of Element 4 considerations in the future.

Element 2: Precisely Defined Constructs

Think aloud methods detected design issues for three items: Grade 4, items 3 and 6, and Grade 8, item 3. According to the data, each of the items were challenging to students because of an unclear construct. Among the students who answered Grade 4, item 3 incorrectly were one student with an impairment, one ELL, one student with a cognitive disability and one English proficient student without a disability. Each of these students failed to attend to or interpret a message that indicated that the item required a two-part solution. Because the construct was unclear to these students, they applied their mathematics skills incorrectly.

Likewise, Grade 4 item 6 required students to use a map. What types of calculations were necessary were not clear to students in this item. The assumed construct being testing for this item was the ability to engage in addition in an authentic context. However the context selected introduced a variety of construct-irrelevant variance (Haladyna, Downing, & Rodriguez, 2002) that was identified by student utterances, including confusion about which numbers to add and confusion about how the map worked (arrows were used in two different ways on the surface of the map).

Finally, analysis of Grade 8, item 3 revealed that eight students who provided incorrect answers were confused or distracted by aspects of the item. Four of these students were distracted by the illustration, and four misunderstood what the item was asking. Of the students who were confused or distracted by the illustrations, two students were confused by the content-irrelevant illustrations found on the page of the item. One student was also confused by content-relevant graphic information, not understanding how a relevant illustration related to the item.

Element 3: Accessible, Non-biased Items

Thompson et al.'s Element 3 addresses the accessibility of items and cautions against bias that may be found in particular items. In this study, think aloud methods detected issues related to bias in Grade 8, items 4 and 6.

Sixteen of 43 participants answered Grade 8, item 3 incorrectly. For four of these students, failure to correctly respond to this item was likely due to inexperience with this particular type of

item. One student with a learning disability, 2 ELLs and one English proficient student without a disability became flustered with an item-type that they had never been exposed to and in the process calculated the wrong numbers, incorrectly estimated, or incorrectly guessed.

Results from Grade 8, item 6 revealed possible experiential bias for Grade 8, item 3. Among the 36 students who incorrectly answered this item, two students (one student with a learning disability and one ELL) were confused by the experiential requirements of the item, that is, the item required that students know that a pair of objects mentioned were separate items, not part of a set. Students who answered this item presumed that the objects belonged to a set based on their personal experiences.

Element 5: Simple, Clear, and Intuitive Instructions

Five of the 12 items selected for this research appeared to have issues related to simple, clear, or intuitive instructions (Grade 4, items 2, 3, and 5; Grade 8, items 2 and 3). Grade 4, item 2 was answered correctly by only 16 of 40 participants. Student responses to the item led researchers to believe that students failed to recognize the two-part nature of this item.

Ten out of 39 students in the sample answered Grade 4, item 3 incorrectly. The content of the instructions confused a number of students. Five of the ten students who selected an incorrect answer believed the item instructions asked students to order a number of objects in the exact opposite order specified by the instructions (English proficient student without a disability = 1, ELL = 2, Deaf/Hard of Hearing = 1, Cognitive Disability = 1).

Only 7 out of 40 students answered Grade 4, item 5 correctly. In this item, the placement and nature of the directions influenced how students approached the problem (instructions were given at the top and bottom of the page, with a workspace in the middle). Four students who were confused did not see that the item instructions were divided by a workspace (3 ELLs and 1 Deaf/HH student). These students read the portion of the directions that was above the workspace and stopped.

Twenty out of 43 students in the sample answered Grade 8, item 1 correctly. Ten students (2 English proficient students without disabilities, 3 ELLs, 2 Deaf , and 3 students with a learning disability) answered incorrectly because of unclear instructions and procedures. Think aloud data revealed that this item did not have clear and intuitive instructions and procedures when students confused why a particular portion of the item was emphasized in text. These students interpreted the emphasis of an italicized word in a variety of ways, all incorrect.

During protocols for Grade 8, item 3, four students misinterpreted the meaning of what the item was asking (1 LD, 2 ELLs, and 1 English proficient student without a disability). The four

students believed that the instructions directed them to do double the work of what test designers had intended. Consequently, all answered incorrectly.

Element 6: Maximum Readability and Comprehensibility

Think aloud methods detected issues related to readability or comprehensibility for four items (Grade 4, item 3; Grade 8, items 1 and 2). Most students (29 of 39) were successful on Grade 4, item 3. Those who were not successful revealed comprehensibility issues related to the text in the item. In addition to the students who were confused by the construct tested (see above), two more English proficient students without disabilities were unfamiliar with the key word in the item. One student crossed out the key word and replaced it with a word she thought to be a synonym.

Only 18 out of 42 students answered Grade 8, item 1 correctly. The reason, as apparent from think aloud data, was most likely the vocabulary in the item. Many students had difficulty sounding out both the vocabulary in the item and several never fluently read the two words. However, these students still appeared to understand that the item referred to two designs located directly above the item. It is unknown whether replacing the challenging vocabulary with other mathematical terms would have changed the construct.

Grade 8, item 2 was also challenging to students. As noted above, only 6 out of 42 students answered this item correctly. Aside from issues related to unfamiliar non-mathematical vocabulary (Element 3), one student with a hearing impairment also confused a term that had a double meaning.

Element 7: Maximum Legibility

Finally, think aloud methods detected legibility issues in two items (Grade 4, item 6; Grade 8, item 6). As noted above, only 10 of 39 students answered Grade 4, item 6 correctly. Five students who answered incorrectly misread the numbers on the page. Because of the font selected for the item, two numbers ("1" and "7") looked remarkably similar. This similarity confused five students, who miscopied information from the item, but otherwise demonstrated the ability to perform addition operations.

Grade 8, item 6 caused problems for four students (ELL = 2, English proficient student without a disability = 1, LD = 1). The cause for the misreading was most likely an editing error. For one fraction, an automatic function on the computer aligned a fraction in one way (e.g., 1 ½) but did not align the second fraction (e.g., 7 7/8). Thus, the confusing nature of the second fraction mislead student into performing incorrect operations.

Summary

Tables 4 and 5 show the results and conclusions from the think aloud analyses for grades 4 and 8. Think aloud methods were useful in detecting problems related to universal design elements for nine items. They were also successful in determining that construct-relevant content was too challenging for students on one item.

Table 4. 4th Grade Item Results and Conclusions

Item	Results	Conclusions
4th Grade, Item 1	Content too difficult for students.	
4th Grade, Item 2	Students were confused or did not understand how the word "more" associated with correct answer.	Detected Element 5 (simple, clear and intuitive instructions and procedures) deficiencies, but demonstrated that Element 7 (maximum legibility) may not be enough to facilitate student understanding.
4th Grade, Item 3	Students appeared unclear on words "most" and "fewest."	Detected deficiencies in Element 2 (precisely defined constructs), Element 5 (simple, clear and intuitive instructions and procedures), and Element 6 (maximum readability and comprehensibility).
4th Grade, Item 4	Data inconclusive as to why students had difficulties.	Think aloud data cannot detect design deficiencies for all items.
4th Grade, Item 5	Instructions confusing to some students. Think aloud data may have confounded confusion.	Detected deficiencies in Element 5 (simple, clear and intuitive instructions and procedures). Think aloud methods should be carefully approached in items that ask students to "explain."
4th Grade, Item 6	Students misread map.	Detected deficiencies in Element 2 (precisely defined constructs) and Element 7 (maximum legibility).

Table 5. 8th Grade Item Results and Conclusions

Item	Results	Conclusions
8th Grade, Item 1	Language in instructions confused students.	Detected deficiencies in Element 5 (simple, clear and intuitive instructions and procedures) and Element 6 (maximum readability and comprehensibility).
8th Grade, Item 2	Students confused by instructions asking them to pick "best" choice.	Detected deficiencies in Element 5 (simple, clear and intuitive instructions and procedures) and Element 6 (maximum readability and comprehensibility).
8th Grade, Item 3	Construct unclear to some students.	Detected deficiencies in Element 2 (precisely defined constructs) and Element 5 (simple, clear, and intuitive instructions).
8th Grade, Item 4	Students unable to answer item correctly because they were unfamiliar with content.	Think aloud method was effective in demonstrating opportunity to learn (OTL) discrepancies between students. Detected Element 3 (accessible, non-biased items) deficiencies.

Table 5. 8th Grade Item Results and Conclusions (continued)

Item	Results	Conclusions
8th Grade, Item 5	Data inconclusive as to why students had difficulties.	Think aloud data cannot detect design deficiencies for all items.
8th Grade, Item 6	Language used in item confusing or unfamiliar to students.	Detected deficiencies in Element 3 (accessible, non-biased items), Element 6 (maximum readability and comprehensibility), and Element 7 (maximum legibility).

Data were inconclusive for two items. Grade 4, item 4 and Grade 8, item 5 were challenging to students, but think aloud methods data did not provide us with any information on why students struggled. Of the 27 students who selected the incorrect response for Grade 4, item 4, 23 of 27 were not distracted or confused by design aspects of the item. The content irrelevant picture along the right side of the page only distracted one student. Three students did not provide enough information to determine distraction/confusion. The rest of the incorrect answers are likely attributable to students not understanding item content, but not being able to express their thoughts. For this item, students receiving ELL services appeared to have had more difficulty identifying and engaging in correct problem-solving strategies than their English language proficient peers. Based on the methods they used to solve the item, it is probable that the students did not fully understand the mathematics that the item required.

Similar results occurred for Grade 8, item 5. Although this item contained a considerable amount of text, most students were able to identify the key information. Specifically, 32 of the 42 identified the equation, although only 7 were able to successfully solve the equation. Six of these students, however, did not provide enough information to determine whether they were distracted or simply did not understand the content. Consequently, it is not possible to determine the extent to which this item contained unnecessarily confusing or distracting design.

Grade 4, item 4 and Grade 8, item 5 demonstrated that think aloud methods are not always adequate for discovering design issues in large-scale assessments. Such data support Leighton's (2004) critique of think aloud methods: that the method is ineffective for very difficult items largely because highly-skilled students work automatically and less-skilled students have trouble explaining why they do not understand. The mathematics level of Grade 4, item 4 and Grade 8, item 5 was obviously very high, and challenging to students. Because of this, many students were unable to explain why they struggled. Without obvious design issues contributing to student difficulty, it is hard to determine the exact source of student miscue.

Table 6 summarizes information by correct or incorrect answer and the source of challenge for subgroups. The percentages in the table indicate the percentage of students whose think aloud information indicated that an incorrect answer was due to the item's illustrations or wording,

that there was no reason to believe that design caused issues, or that the reason for an incorrect answer was unclear. These data demonstrated that, overall, more students were not affected by design issues than were. There were, however, several students of all categories who were challenged by item design issues. Issues were found relative to constructs, accessible/non-biased items, unclear instructions, incomprehensible language, and illegible text.

Table 6. Item Summary Information

	Answer	Distraction			
		Illustrations	**Wording**	**No**	**Not Apparent**
	Incorrect				
LD		6 (5.56%)	15 (13.89%)	34 (31.48%)	7 (6.48%)
Deaf/HH		8 (6.40%)	15 (12.00%)	39 (31.20%)	13 (10.40%)
CogD		2 (4.87%)	4 (9.76%)	10 (24.39%)	24 (58.54%)
ELL		7 (7.37%)	16 (16.84%)	28 (29.47%)	19 (20.00%)
Non-D		6 (4.88%	10 (8.13%	30 (24.39%)	9 (7.32%)
	Correct				
LD			4 (3.70%)	39 (36.11%)	3 (2.78%)
Deaf/HH			1 (0.80%)	47 (37.60%)	2 (1.60%)
CogD				1 (2.44%)	
ELL		1 (1.05%)	1 (1.05%)	20 (21.05%)	3 (3.16%)
Non-D			1 (0.08%)	65 (52.85%)	2 (1.63%)

Most fourth-grade and eighth-grade students with disabilities were able to verbalize while thinking aloud, which seems contrary to earlier findings with young students (Branch, 2000). Students with learning disabilities were very capable of "thinking aloud." So too were ELLs, although translator services may be necessary for students with very low levels English proficiency. Students who were deaf and hard of hearing were able to think aloud with the assistance of sign language translators.

 Students with cognitive disabilities, however, had the greatest difficulty producing both introspective and retrospective data. Think aloud data for this sub-sample is questionable and was used sparingly for analysis. Most students with significant cognitive disabilities participate in state "alternate" assessments. As is noted from the sample for this study, only nine students with significant cognitive disabilities could be found in an entire school district that took the general grade level assessment. Of these, one student was taken from a grade level above, simply to add numbers to the sample. The nine students who participated in the study were not able to provide succinct information verbally. For example, students often had trouble understanding what was required of them on items and frequently did not have the skills to approach solving problems.

Despite the challenges faced by students with cognitive disabilities, these students are increasingly included in statewide assessments and should be included in think aloud studies as much as possible. Future research is needed to develop a think aloud method for students with cognitive disabilities that aids in understanding the problem-solving processes of this subgroup. One method that may be helpful is training teachers in protocol analysis. Fuchs and Fuchs (1989) found that students with disabilities consistently perform better on a range of assessments when in the presence of familiar examiners. If "thinking aloud" is considered a performance activity, it is logical to assume that students might be more proficient (and more comfortable) around familiar research assistants.

A second technique that may aid students with cognitive disabilities to "think aloud" is to change the think aloud protocol altogether. One possibility for a different protocol is to adapt Van Someren et al.'s (1994) method of asking students to act like a teacher and instruct the researcher in how to solve a problem. This style is more interactive than traditional think aloud protocols, and may be a better method of gathering information from this population. Combined with a familiar research assistant, more valid results may be yielded from students with cognitive disabilities in the future. More valid research results may in turn provide important information on the validity of large-scale assessments overall for students with cognitive disabilities.

Future Directions

The think aloud method (and subsequent analyses) appears to provide important information on test design issues. By understanding the process that students use to solve problems on large-scale tests, we can more easily determine what design features may interfere with effective problem-solving. As the emphasis on the information derived from testing increases with each passing year, so does the importance of understanding the processes students use to solve problems on tests (and design-related issues related to effective student processing).

Along with changes that should be made for students with cognitive disabilities, a second future direction for think aloud research is to continue targeting other subgroups mentioned in the No Child Left Behind Act (Kopriva, 2001). By using statewide data, researchers and states can determine which groups are perennially underperforming on statewide tests and use think aloud protocols to better understand processes these subgroups use to solve problems. Replication of the above-mentioned methods may be effective for most subgroups, and may provide many needed answers about why students struggle or succeed on tests.

This research demonstrated that design issues can be detected in tests when students think out loud while they are solving problems. The design issues detected by students themselves are

important to note in order to increase the validity of test results. A variety of methods can be used to improve the test design process. One method that was found to be particularly effective in this study was the think aloud method.

References

Branch, J. L. (2000). Investigating the information-seeking processes of adolescents: The value of using think-alouds and think afters. *Library and Information Science Research, 22*(4), 371–392.

Bogdan, R. C., & Biklen, S. K. (1992). *Qualitative research for education. An introduction to theory and methods*. Boston: Allyn and Bacon.

Camburn, E., Correnti, R., & Taylor, J. (2000). *Using qualitative techniques to assess the validity of teachers' responses to survey items*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 24–28, 2000.

Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment, 3*(7). Retrieved from http://www.bc.edu/research/intasc/jtla/journal/v3n7.shtml

Duncker, K. (1945). On problem-solving. In Dashiell, J. F. (Ed.) *Psychological Monographs* (pp.1–114). Washington, DC: American Psychological Association.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (Revised edition)*. Cambridge, MA: MIT Press.

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative Health Research, 3*(4), 430–441.

Fuchs, L., & Fuchs, D. (1989). Effects of examiner familiarity on Black, Caucasian, and Hispanic children: A meta-analysis, *Exceptional Children, 55*, 303–308.

Grise, P., Beattie, S., & Algozzine, B. (1982). Assessment of minimum competency in fifth grade learning disabled students: Test modifications make a difference. *Journal of Educational Research, 76*, 35–40.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: National Center on Educational Outcomes.

Ketterlin-Geller, L. R. (2005). Knowing what all students know: Procedures for developing universally designed assessments. *Journal of Technology, Learning, and Assessment. 4*(1). Retrieved from http://www.bc.edu/research/intasc/jtla/journal/pdf/v4n2_jtla.pdf

Kopriva, R. (2001). *ELL validity research designs for state academic assessments: An outline of five research designs evaluating the validity of large-scale assessments for English language learners and other test takers.* Paper prepared at the Council of Chief State School Officers Meeting, Houston, TX, June 22–23, 2001.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 6–15.

Nielson, J. (1994). Estimating the number of subjects needed for a thinking aloud Test. *International Journal of Human-Computer Studies, 41*(3), 385–397.

Nielson, J., Clemmensen, T., & Yssing, C. (2002, October). *Getting access to what goes on in people's heads? Reflections on the think-aloud technique*. NordiCHI, 101–110.

Nolan, M. J., & Chandler, K. (1996). *Use of cognitive laboratories and recorded interviews in the National Household Education Survey*. Technical Report. Washington, DC: National Center for Educational Statistics.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Earlbaum.

Rothman, R. (1993). ACT unveils new assessment, planning system. *Education Week, 12* (24), 1, 17.

Shriver, K. A. (1984). *Revising computer documentation for comprehension: Ten exercises in protocol-aided revision*. CDC Technical Report Number 14. Pittsburgh, PA: Carnegie Mellon University: ERIC Document Reproduction Service Number ED Z78943.

Shriver, K.A. (1991). *Plain language for expert or lay audiences: Designing text using protocol-aided revision*. Technical Report Number 46. California: ERIC Document Reproduction Service Number ED 278943.

Thompson, S. J., Johnstone, C.J., & Thurlow, M.L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: National Center on Educational Outcomes.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think-aloud method: A practical guide to modeling cognitive processes*. San Diego, CA: Academic Press Ltd.