Abstract Title Page

Title: Exploring the utility of student-think alouds for providing insights into students' metacognitive and problem-solving processes during assessment development

Authors and Affiliations:

Deni Basaraba, Ph.D. (Southern Methodist University)
Yetunde Zannou, Ph.D. (Southern Methodist University)
Dawn Woods, M.A. (Southern Methodist University)
Leanne Ketterlin-Geller, Ph.D. (Southern Methodist University)

Background/Context

Within the current test-centered educational reform movement, considerable emphasis is placed on using assessment results to make instructional decisions for individual students. Test scores are used to estimate a student's current level of skill, monitor his or her progress during instruction, and identify whether the student has gained the expected skills and knowledge at the end of an instructional period. Informed and appropriate instructional decision-making, however, is predicated on several assumptions: (a) assessments can provide information about students' understanding, strategic competence, and reasoning skills within a domain and (b) that this information can be obtained accurately and efficiently to help educators determine whether students will benefit from additional instructional supports to be successful (Ikeda, Neessen, & Witt, 2008). Meeting these assumptions, however, requires taking steps during the assessment development process to ensure valid inferences can be made about students' knowledge, skills, and abilities based on their performance on the assessment.

Verbal Protocols: Insights Into Students' Thinking: According to the National Center on Educational Outcomes (NCEO) and others (Ericsson & Simon, 1993; Gorin, 2007), verbal protocols, or "think-alouds" are effective verification and validation tools during item and assessment development (Johnstone, Bottsford-Miller, & Thompson, 2006). Two methods are available for collecting student think-aloud data: concurrent and retrospective think-alouds (Ericsson & Simon, 1993). During a concurrent-think aloud, students are instructed to think aloud as they complete a task with the intent of unveiling the cognitive processes they engage in and the information they attend to *while* they are solving a problem (Leighton & Gierl, 2007). In contrast, retrospective think-alouds are conducted *after* students have solved the problem, providing students an opportunity to describe any metacognitive processes or higher-level problem-solving processes they may have engaged in while solving the problem that were omitted during the concurrent think-aloud (Ericsson & Simon, 1993). Retrospective think-alouds also provide an opportunity for assessment developers to obtain valuable student input about specific characteristics of the assessment items being developed, such as their understanding of the task and whether the content, language, or structure of the problem presented any challenges (Almond et al., 1999).

Metacognition: Thinking About Thinking. Metacognition refers to one's knowledge and deeper understanding of the cognitive processes and products one engages in while completing a task (Flavell, 1979) and is a skill frequently assessed using verbal protocols given their intent to make overt the covert cognitive processes one engages in while completing a task (Fonteyn, Kulpers, & Grobe, 1993; Rosenzweig, Krawec, & Montague, 2011). Research indicates that awareness of students' metacognition is useful to provide a better understanding of the factors that contribute to successful mathematics performance (Desoete, Roeyers, & Buysse, 2001). Predictive, planning, and procedural metacognitive skills are particularly relevant because these types of metacognition occur during initial problem solving through the final stages of problem solving when students check the outcomes of their work and interpret the results (Desoete et al., 2001). Prediction in mathematics refers to the ability to accurately distinguish challenging tasks from easier tasks and using that information to help students concentrate their efforts on those tasks requiring more effort. Planning refers to the ability to analyze a given task, to retrieve relevant domain-specific knowledge or skills needed and to sequence problem-solving strategies needed to complete the task. Procedural metacognition refers to awareness of one's thinking processes and the ability to accurately explain how one solved a problem or completed a task (Jacobs & Paris, 1987). Previous research (Desoete et al., 2001) indicates that 38% of the variance in students' off-line metacognition behaviors (e.g., predictive, planning, and procedural metacognition) was explained by their mathematics ability group, with above-average mathematical problem-solvers performing better on researcher-developed scales of off-line metacognition.

Study Purpose & Research Questions

The purpose of this study is to describe the development of a universal screening assessment of algebra readiness for Grades 2 -4 with a particular focus on the collection of student think-aloud data to (a) provide insights into students' cognitive and problem-solving processes, (b) investigate the relationships between students' predictive, planning, and procedural metacognition and their performance on multiple-choice mathematics items that were similar in structure and content to those that would be used in the universal screening assessment and (c) inform the item writing and revision process. We explore whether sample items written for the verbal protocol data collection required students to use varying levels of cognitive processing that align with research-based intertwined strands of mathematical proficiency (Kilpatrick, Swafford & Findell, 2001), whether there were any specific item features that interfered with our ability to adequately measure the constructs of interest, and explain how information obtained from student think-alouds can inform the item writing and revision processes during assessment development. Our specific research questions are: (a) To what extent can multiple-choice mathematics items be written that require students' to demonstrate varying levels of cognitive engagement; and (b) What is the relation between students' predictive, planning, and procedural metacognition and their performance on multiple-choice mathematics items. We will also discuss how information obtained from student think-alouds be used to inform the item development process and report the preliminary findings of a large-scale pilot study with approximately 20,000 students in Grades 2-5 (students from Grade 5 are being included due to the recent adoption of revised content standards in which some content is new to Grade 4 but will have been seen by current Grade 5 students).

Setting & Participants

The think-aloud study was conducted in one elementary school in the Southwest. Trained research members interviewed 30 students in Grades 2-4 (10 students per grade level) who were identified by their classroom teachers ($N = 6$) on classroom grades and curriculum based measures (CBMs) of mathematics as having low, moderate, or high mathematics ability. The 10 second grade students ranged in age from 7-9 years, 50% were female, and 70% were Caucasian. The 10 third grade students ranged in age from 8-9 years, 40% were male, and 80% were Caucasian. The 10 fourth grade students ranged in age from 9-10 years, 70% were female, and 90% were Caucasian. Information for students participating in the pilot test is not yet available.

Intervention/Program/Practice

Student think-aloud data collected as part of this study were collected during the development of the Elementary School students in Texas Algebra Ready (ESTAR) universal screening measure for Grades 2-4. Similar to other universal screening assessment systems, the ESTAR Universal Screener will include three equivalent, alternate forms of approximately 24 items each to be administered to *all* students in Grades 2-4 in the fall, winter, and spring to help teachers identify students who may need additional instructional supports in mathematics. Items focus on key algebra-related content and were purposefully written to align with state content standards.

Items were also written to represent four levels of cognitive engagement (procedural fluency, conceptual understanding, strategic competence, and adaptive reasoning). Items assessing *procedural fluency* use formal language or symbolic representation to measure a students' ability to accurately carry out computations, follow multiple steps sequentially, and/or use algorithms and knowledge of mathematical properties to solve mathematical problems. Items assessing *conceptual understanding* are intended to provide students with an opportunity to demonstrate an integrated and functional grasp of mathematical ideas by (a) understanding a specific task as it relates to a larger concept, (b) finding relationships between pieces of information, (c) making connections between similar representations, and/or (d) using models and multiple representations to solve problems. Items assessing *strategic competence* assess the ability to formulate a problem in mathematical terms, to represent problem-solving strategically (e.g., verbally, symbolically, graphically, etc.), and/or to identify and effectively use an appropriate strategy to solve a problem. Finally, items assessing *adaptive reasoning* assess a students' ability to think logically about a problem and deductively select an appropriate problem-solving approach, to rationalize and justify the strategy used to solve a problem, and/or to appropriately explain a procedure or concept.

Research Design

The verbal protocol data were collected with a convenience sample of 30 students in Grades 2-4 in a local elementary school. All students within each grade solved the same 10 multiple-choice math problems and answered the same questions during the retrospective think-aloud. The pilot test data will be collected with a sample of approximately 20,000 students in schools and districts that volunteered to participate. Care was taken to stratify the items to ensure adequate representation of the content standards, levels of cognitive engagement, and levels of difficulty; students were randomly assigned to grade-level forms.

Data Collection & Analysis

In this section we describe the data collection and analyses for both components of the proposal: (a) verbal protocol data collection, and (b) universal screener pilot test.

Verbal protocols. Think-aloud interviews were conducted with 30 students in Grades 2-4 by eight trained interviewers over the course of two weeks. Each student responded to 10 multiple-choice items during the interview; we refer to these items as "item shells" with the intent that, if students' response indicated that the items were time-efficient and assessed the targeted construct and level of cognitive engagement that very similar items using the same format could be written for the pool of items that would be used to develop the universal screener. Parental permission and student assent were obtained for all participants and each interview, which, when divided into two separate sections to minimize student fatigue, lasted anywhere from 45-60 minutes and were videotaped   All recordings were transcribed to ensure that we accurately captured students' language during the interviews. As noted earlier, during the interview we asked students to complete a concurrent think aloud while solving each problem and then asked a series of targeted questions during the retrospective think-aloud; these questions are available in Appendix B.

Data obtained from these interviews will be examined descriptively as we investigate the number and percent of students who selected the correct response and each distractor and whether the levels of cognitive engagement and relative difficulty targeted by each item were correct. In addition, we calculate descriptive statistics of students' response times for each item; because the ESTAR Universal Screener is intended to be time and resource-efficient we wanted to ensure that some items weren't so challenging that they required an extensive amount of time for students to solve. Student think-alouds for each item will also be coded with respect to challenges posed by the language, vocabulary and graphics associated with item. To examine our research questions about the relations between types of metacognition and mathematical problem-solving we will conduct three analyses. First, we will conduct Spearman Rho correlations to account for our continuous dependent variable (mathematical problem-solving scores) and categorical independent variables (ratings of students' predictive, planning, and procedural metacognition). Second, to investigate the relation between types of metacognition and general mathematical problem solving ability using multiple regression to see whether the three types of metacognition are equally predictive of students' math problem solving skills. Finally, to explore the possible interaction between the types of metacognition, relative difficulty, and mathematical content we will calculate odds ratios to examine whether any relations between the three variables are due to something other than chance.

Pilot study. Item-level data from a convenience sample of approximately 20,000 students enrolled in 83 schools in 50 districts. All participating students will take one of 26 alternate forms of the computer-based assessment. Each form is comprised of 19 unique and 5 anchor items (to support the later equating of forms) and the items for each form were purposefully stratified to ensure a relatively equal distribution of content standards, level of cognitive engagement, and relative difficulty. Data obtained from these assessments will be analyzed using IRT modeling to determine the empirical item difficulties. These item difficulties will be examined to determine whether there are consistent differences in difficulty among items measuring the four different levels of cognitive engagement.

Results

Results from the pilot test are currently not available given that all items will be pilot tested in schools this spring but will be presented at the conference. Preliminary results from the verbal protocol data analysis are reported below.

Verbal protocols. Descriptive information regarding the distribution of students' responses for each item, the level of cognitive engagement and relative difficulty, and the average response time for each item is reported, by grade level, in Appendix B. The number of students who selected the correct response, in combination with students' average response times for items and language from their retrospective think-alouds prompted us to think carefully about revisions to one item shell for Grade 3, and five item shells in Grade 4 with respect to level of cognitive engagement. Students' responses during the think-aloud provided even more valuable insights with respect to the relative difficulty of the item shells, however, as evidenced by the fact that we were compelled to re-classify the difficulty of two items for Grade 2, three items for Grade 3, and four items for Grade 4 (30% of all items). In addition, students' responses to the retrospective think-aloud questions provided valuable insights into some of the challenges posed by the language, vocabulary, and graphics associated with the items; we will discuss how we incorporated this feedback into item revisions and future item development.

Conclusions:

Descriptive analyses of the verbal protocol data reveal, similar to other studies, that having students think-aloud while engaging in problem-solving tasks can provide valuable insights into their cognitive processes (Rosenzweig et al., 2011) that can subsequently be used to inform future item development (Almond et al., 2009) and be used to validate the construct(s) being measured (Embretson & Gorin 2001). In addition, examination of the degree to which the levels of cognitive engagement and relative difficulty were verified indicates that items can be purposefully be written to increase in cognitive complexity, which is important given the need for students to be able to conceptually understand and think deeply about mathematics problems beyond rote memorization and procedural fluency (Kilpatrick et al., 2001).

Appendices

Appendix A. References

Almond, P. J., Cameto, R., Johnstone, C. J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C. E., Roach, A. T., & Sato, E (2009). *White paper: Cognitive interview methods in reading test design and development for alternate assessments based on modified academic achievement standards (AA-MAS).* Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

Desoete, A., Roeyers, H., & Buysse, A. (2001). Metacognition and mathematical problem solving in grade 3. *Journal of Learning Disabilities, 34,* 435-449.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38,* 343-368.

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

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34,* 906-911.

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

Gorin, J. S. (2007). Test construction and diagnostic testing. In J. P. Leighton & M. J. Gierl (Eds.) *Cognitive diagnostic assessment for education: Theory and application* (pp. 173-204). New York, NY: Cambridge

Ikeda, M. J., Neessen, E., & Witt, J. C. (2008). Best practices in universal screening. In A. Thomas & J. Grimes (Eds.) *Best practices in school psychology* (Vol. 2, pp. 103-114). Bethesda, MD: National Association of School Psychologists.

Jacobs, J. E., & Paris, S. G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, 22,* 255-278.

Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved March 15, 2012.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academy Press.

Leighton, J. P., & Gierl, M. J. (2007). Verbal reports as data for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.) *Cognitive diagnostic assessment for education: Theory and application* (pp. 146-172). New York, NY: Cambridge.

Rosenzweig, C., Krawec, J., & Montague, M. (2011). Strategy use of eighth-grade students with and without learning disabilities during mathematical problem-solving: A think-aloud analysis. *Journal of Learning Disabilities, 44,* 508-520.

Appendix B. Tables and Figures

Figure 1.
Directions for Concurrent and Retrospective Think-Alouds

Directions for Concurrent Think-Aloud:

I am going to ask you to solve some math problems and to talk about how you solved the problems, just like you do in class. We are interested in understanding the thinking you use while solving math problems. Today, I want you to say **all** of your thoughts about how you came to your answer out loud, rather than thinking them in your head.

If you forget to tell me what you are thinking while solving the problems I will remind you to do so by asking you to "Keep talking." When you are done working on each problem I will ask you some questions about what you thought about the problem.

Questions Asked During Retrospective Think-Aloud

1. **You rated this question as** [very hard, hard, easy, very easy]. **Why did you rate this question as** [very hard, hard, easy, very easy]?
2. **What do you know about _________?** [We provided the target skill for the item, such as adding two digit numbers, multi-step word problems, etc.]
3. **What is this problem asking you to do?**
   [Prompt: What question are you answering?]
4. **What information do you need to solve the problem?**
   [Prompt: What words or numbers helped you solve the problem?]
5. **What strategies and steps did you take to solve the problem?**
   [Prompt: What did you do first? What did you do second? etc.]
6. **Does your answer for this problem make sense? Why?**
   [Prompt: How do you know?]
7. **Was there any part of the question that was confusing or hard to understand?**
   [If yes, ask the student to describe the specific part(s) of the problem that were confusing or hard to understand]
8. **Were there any words that were hard to understand?**
   [If yes, ask the student to describe the specific aspects of the language such as vocabulary or sentence length that were confusing or hard to understand]
9. **Explain why you chose this answer over all of the other options.**
10. **Explain why you did not choose the other answer options.**
    [Asked students to articulate why they didn't choose the three other non-selected responses]

Table 1
Descriptive Analyses of Grade 2 Verbal Protocol Data

Item Number	Targeted Level of		Verified Level of		Number and Percent of Students Who Selected				Average Response Time
	Cognitive Engagement	Difficulty	Cognitive Engagement	Difficulty	Response A	Response B	Response C	Response D	
1	Procedural	Easy	Yes	Yes	0% (0)	22.2% (2)	**66.7%** **(6)**	11.1% (1)	2:10
2	Strategic	Difficult	Yes	Yes	0% (0)	**90%** **(9)**	0% (0)	10% (1)	1:25
3	Conceptual	Easy	Yes	Yes	0% (0)	0% (0)	70% (7)	**30%** **(3)**	1:15
4	Conceptual	Medium	Yes	No	**10%** **(1)**	10% (1)	40% (4)	40% (4)	1:49
5	Procedural	Easy	Yes	Yes	**100%** **(10)**	0% (0)	0% (0)	0% (0)	0:44
6	Procedural	Medium	Yes	Yes	0% (0)	0% (0)	28.5% (2)	**71.4%** **(5)**	2:21
7	Conceptual	Medium	Yes	Yes	11.1% (1)	0% (0)	**88.9%** **(8)**	0% (0)	0:56
8	Procedural	Difficult	Yes	No	10% (1)	**70%** **(7)**	20% (2)	0% (0)	1:24
9	Adaptive	Medium	Yes	Yes	0% (0)	60% (6)	**40%** **(4)**	0% (0)	1:21
10	Procedural	Easy	Yes	Yes	0% (0)	**80%** **(8)**	20% (2)	0% (0)	0:47

Note: Values for number and percent of students who selected the correct response are in bold text.

Table 2
Descriptive Analyses of Grade 3 Verbal Protocol Data

Item Number	Targeted Level of		Verified Level of		Number and Percent of Students Who Selected				Average Response Time
	Cognitive Engagement	Difficulty	Cognitive Engagement	Difficulty	Response A	Response B	Response C	Response D	
1	Procedural	Medium	Yes	Yes	**70%** **(7)**	10% (1)	0% (0)	20% (2)	2:06
2	Conceptual	Medium	Yes	Yes	20% (2)	60% (6)	**20%** **(2)**	0% (0)	2:03
3	Strategic	Difficult	Yes	Yes	40% (4)	0% (0)	10% (1)	**50%** **(5)**	2:09
4	Conceptual	Easy	Yes	No	30% (3)	**40%** **(4)**	30% (3)	0% (0)	4:32
5	Conceptual	Easy	Yes	Yes	**70%** **(7)**	20% (2)	10% (1)	0% (0)	2:02
6	Strategic	Difficult	Yes	Yes	10% (1)	30% (3)	40% (4)	**20%** **(2)**	2:20
7	Adaptive	Difficult	No	No	0% (0)	0% (0)	**100%** **(10)**	0% (0)	1:05
8	Conceptual	Easy	Yes	Yes	**80%** **(8)**	0% (0)	0% (0)	20% (2)	0:58
9	Procedural	Easy	Yes	Yes	0% (0)	**100%** **(10)**	0% (0)	0% (0)	0:28
10	Procedural	Medium	Yes	No	10% (1)	10% (1)	10% (1)	**70%** **(7)**	1:25

Note: Values for number and percent of students who selected the correct response are in bold text.

Table 3
Descriptive Analyses of Grade 4 Verbal Protocol Data

Item Number	Targeted Level of		Verified Level of		Number and Percent of Students Who Selected				Average Response Time
	Cognitive Engagement	Difficulty	Cognitive Engagement	Difficulty	Response A	Response B	Response C	Response D	
1	Conceptual	Medium	Yes	Yes	10% (1)	10% (1)	**70%** **(7)**	10% (1)	2:14
2	Procedural	Easy	Yes	Yes	10% (1)	0% (0)	0% (0)	**90%** **(9)**	2:53
3	Strategic	Easy	No	Yes	0% (0)	**100%** **(10)**	0% (0)	0% (0)	2:08
4	Procedural	Medium	No	No	**30%** **(3)**	10% (1)	0% (0)	60% (6)	2:09
5	Conceptual	Medium	No	Yes	0% (0)	**100%** **(10)**	0% (0)	0% (0)	2:28
6	Procedural	Medium	No	Yes	30% (3)	0% (0)	**70%** **(7)**	0% (0)	3:08
7	Conceptual	Medium	Yes	No	**50%** **(5)**	0% (0)	40% (4)	10% (1)	4:04
8	Procedural	Difficult	No	Yes	0% (0)	12.5% (1)	25% (2)	**62.5%** **(5)**	5:17
9	Adaptive	Medium	No	Yes	0% (0)	0% (0)	**100%** **(10)**	0% (0)	2:15
10	Adaptive	Medium	Yes	No	**66.7%** **(6)**	33.3% (3)	0% (0)	0% (0)	4:20

Note: Values for number and percent of students who selected the correct response are in bold text.