Applied Measurement in Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Gathering Response Process Data for a Problem-Solving Measure through Whole-Class Think Alouds

Jonathan David Bostic , Toni A. Sondergeld , Gabriel Matney , Gregory Stone & Tiara Hicks

To cite this article: Jonathan David Bostic , Toni A. Sondergeld , Gabriel Matney , Gregory Stone & Tiara Hicks (2021) Gathering Response Process Data for a Problem-Solving Measure through Whole-Class Think Alouds, Applied Measurement in Education, 34:1, 46-60, DOI: 10.1080/08957347.2020.1835913

To link to this article: https://doi.org/10.1080/08957347.2020.1835913

Published online: 01 Feb 2021.

APPLIED MEASUREMENT IN EDUCATION
2021, VOL. 34, NO. 1, 46–60
https://doi.org/10.1080/08957347.2020.1835913

Routledge
Taylor & Francis Group

Gathering Response Process Data for a Problem-Solving Measure through Whole-Class Think Alouds

Jonathan David Bostic[a], Toni A. Sondergeld[b], Gabriel Matney[a], Gregory Stone[c], and Tiara Hicks[d]

[a]School of Teaching & Learning, Bowling Green State University, Bowling Green, OH, USA; [b]School of Education, Drexel University, Philadelphia, PA, USA; [c]MetriKs Amérique LLC, OH, Sylvania, USA; [d]George I Sánchez Community Collaborative School, Albuquerque, NM, USA

ABSTRACT

Response process validity evidence provides a window into a respondent's cognitive processing. The purpose of this study is to describe a new data collection tool called a whole-class think aloud (WCTA). This work is performed as part of test development for a series of problem-solving measures to be used in elementary and middle grades. Data from third-grade students were collected in a 1–1 think-aloud setting and compared to data from similar students as part of WCTAs. Findings indicated that students performed similarly on the items when the two think-aloud settings were compared. Respondents also needed less encouragement to share ideas aloud during the WCTA compared to the 1–1 think aloud. They also communicated feeling more comfortable in the WCTA setting compared to the 1–1 think aloud. Drawing the findings together, WCTAs functioned as well if not better, than 1–1 think alouds for the purpose of contextualizing third-grade students' cognitive processes. Future studies using WCTAs are recommended to explore their limitations and other factors that might impact their success as data gathering tools.

1. Introduction

Validity is central to developing and using high-quality assessments (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). The five sources of validity evidence described by the *Standards for Educational and Psychological Testing* (AERA et al., 2014) are test content, response processes, relations to other variables, internal structure, and consequences from testing/bias. Necessary and appropriate evidence related to these sources are needed to generate a validity argument (AERA et al., 2014; Kane, 2013, 2016). Table 1 provides a brief description of the five sources. Evidence for every source is not necessarily required; however, claims should be appropriately drawn from evidence sources, and more validity sources used allows for stronger arguments (AERA et al., 2014; Bostic, Krupa, and Shih, 2019b). In this manuscript, we dig deeper into one source of validity evidence – response process, which indicates the cognitive activities that test-takers employ to think about and answer test items (Hubley & Zumbo, 2017; Leighton, 2017). Our overarching goal is to inform researchers and practitioners about ways to gather response process evidence through think alouds. Specifically, a new approach to think alouds called "whole-class think alouds" or WCTAs is offered and

CONTACT Jonathan David Bostic bosticj@bgsu.edu Bowling Green State University, 529 Education Building, Bowling Green, OH 43403, USA

Table 1. Brief descriptions of the sources of validity evidence.

Source	Validity Evidence Description
Test Content	Determines if the assessment is actually a measure of the intended construct (Kane, 2012) and the extent to which the items include the intended content (AERA et al., 2014). It also takes a deeper look at the questions and compares them to the domains that are presented in state standards. Further, it ensures that the questions are of an appropriate cognitive level and that the questions assess the most important aspects of the domain (Sireci & Faulkner-Bond, 2014).
Response Process	Analyzes how participants might react to the item. It ensures that the interaction between the item and the participant is as desired. This evidence expresses how students engage with the items, but it can also be used to answer questions about why different groups perform differently on the test than others (AERA et al., 2014).
Internal Structure	Evaluates items to determine if they accurately correspond to the intended aspect of the construct (AERA et al., 2014). It also investigates what information the item can provide, check if items introduce bias, and ensures the test is written in a way that is stable.
Relations to Other Variables	Assesses the relationships between the measure of interest and other variables. Evidence can be convergent, meaning that measures of similar things should be related, or discriminant, meaning that measures of different things should be unrelated (AERA et al., 2014).
Consequences of Testing	Investigates possible outcomes that may come from use of the assessment (AERA et al., 2014; Lane, 2014). Consequences may be intended or unintended, and may be positive or negative. This should be explored during test development and again following test use (Lavery et al., 2019).

supported. This study may encourage and inform those who develop, use, and administer assessments for use in K-12 settings.

2. Related Literature

2.1. Building a Shared Language of Validity and Validity Arguments

Research for this study contextualizes validity as “the degree to which evidence and theory support the interpretations of test scores for proposed uses” (AERA et al., 2014, p. 11). Validity, for the purposes within this manuscript, is a unitary construct derived from a validity argument (AERA et al., 2014). The idea of a unitary construct is fairly well agreed upon by many (Plake & Wise, 2014); hence, the authors of this manuscript seek to appropriately and clearly delineate a view on validity that has been readily adopted within the *Standards* (AERA et al., 2014) and sets the tone for the present manuscript. Related to validity is the idea of validation, which is “the process of … accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations” (AERA et al., 2014, p. 11). Thus, an important characteristic of validation scholarship is to gather evidence that helps to support score interpretations for proposed uses. Score interpretations for proposed uses as arguments rest upon a narrative that connects claims and validity evidence in a logical fashion (AERA et al., 2014; Kane, 2013).

Evidence collected during validation studies helps to substantiate a validity argument. A validity argument is developed after conducting validation studies that seek to describe the ways that results and interpretations from instruments might be used (Lavery et al., 2020; Kane, 2012). Validity arguments justify how claims and evidence are connected. They communicate ways in which an instrument and interpretations of its results should and should not be used (AERA et al., 2014; Kane, 2012, 2013, 2016; Lavery et al., 2020). Claims within the validity argument rest upon the quality of the evidence presented, which stem from the methods by which those data are collected. One common data collection tool linked to gathering response process evidence is a think aloud (AERA et al., 2014; Leighton, 2017; Padilla & Benitez, 2014).

2.2. Think Alouds

Think alouds are frequently used in validation studies of assessments used in instructional settings (AERA et al., 2014; Leighton, 2004, 2017). They are designed to gather evidence in support of inferences about how respondents think about the assessment items to which they respond

(Fonteyn, Kuipers, & Grobe, 1993; Leighton, 2004, 2017; Wilson & Miller, 2014). Think alouds, fMRI imaging, and eye-tracking software are all means for gathering response process evidence (AERA et al., 2014; Schindler & Lilienthal, 2019). There is some evidence within mathematics education scholarship suggesting that think alouds provide similar findings about cognitive processes as fMRI and eye-tracking data; albeit, the latter approaches come at a significant cost outside of the scope for many scholars (Schindler & Lilienthal, 2019).

A reason for conducting think alouds in a one-on-one setting (1–1 think alouds) is to gather confirming evidence for a desired cognitive model, often formed from a task analysis (Leighton, 2017). Task analyses suggest knowledge and skills that a respondent might use when engaging with an instrument. Leighton (2017) provides guidance, shared in the following paragraph, about how to conduct think alouds when conducted in a one-on-one setting with a respondent and the context is problem-solving. For the purposes of this study, these types of activities are referred to as 1–1 think alouds.

Traditionally, a respondent and researcher convene in a setting similar to the testing conditions for a 1–1 think aloud (AERA et al., 2014; Leighton, 2017). These 1–1 think alouds are video or audiotaped. They typically begin as a researcher explains the process to a respondent. In some instances, the respondent might engage with a task as practice sharing ideas aloud. After practice, a focal task is administered to the respondent. The researcher generally does not intervene unless the respondent is not readily sharing ideas aloud. It is important that the respondent's ideas come naturally with little intrusion from the researcher, otherwise, those data may be colored because of undue influence. The process continues until all tasks have been completed. Length of time for the 1–1 think aloud is based upon the cognitive complexity of the tasks as well as the developmental readiness of the respondent.

There are many strengths to conducting 1–1 think alouds. First, 1–1 think alouds require relatively little money or expensive technology to complete. Second, they can be easily conducted in a school setting. For example, Bostic and Sondergeld (2015) conducted think-alouds in a school library and meeting room of a school building to gather evidence for a problem-solving measure aligned with grade six Common Core State Standards for Mathematics (CCSSI, 2010). Third, think alouds are not necessarily time intensive. For example, Bostic et al. (2019a) reported that think alouds with elementary students took approximately 3–6 minutes per problem-solving item for a total of 25 minutes per student during a 1–1 think aloud.

On the other hand, there are challenges associated with conducting 1–1 think alouds. First, 1–1 think alouds often remove a student from the classroom for a period of time thus losing instructional time with the classroom teacher. While 25 minutes may not seem excessively long, students may experience some cognitive difficulty readjusting to classwork after missing a segment of classroom instruction. Second, the responses during the 1–1 think alouds may be different from those had the respondent worked with a more familiar person like a classroom teacher. Third, 1–1 think alouds are often conducted outside of a classroom setting. If the items are relevant to content standards as part of classroom instruction, then a classroom teacher who may be interested in seeing the live 1–1 think aloud loses the opportunity to observe and listen. For example, a teacher may want to hear students' thinking on a developmentally appropriate content task with the intention of using those observational data to inform future instruction. Teachers may not necessarily have time to devote to watching 1–1 think aloud videos outside of the school day. Thus, the classroom teacher may not benefit as much from 1–1 think alouds as if they were observed in real time. These challenges present an opportunity to explore a different type of think aloud, which we call a whole-class think aloud (WCTA).

One-on-one think alouds have a place in the literature and have clearly demonstrated their effectiveness in assessing response process validity evidence. As such, they are used in this study as a basis for making comparisons with WCTA data. To that end, this manuscript has two principal aims: (1) describe WCTAs as tools to gather response process data and (2) share strengths and challenges of WCTAs when used as tools to gather validity evidence. The more specific research questions for this study are (a) How do findings from 1–1 and WCTAs compare? and (b) How do 1–1 and WCTAs compare as data analytic tools for gathering response process data?

3. Method

3.1. Contextualizing Items for Study with Think Alouds

Authors for this manuscript have developed a series of problem-solving measures for grades 3–8 (i.e., Problem-solving Measures or PSMs). This study was approved by the Institutional Review Board and is part of a larger grant-funded project (1720646,1720661). All PSMs are composed of mathematics word problems that connect with the Common Core State Standards for Mathematics Content (SMCs) and Standards for Mathematical Practice ([SMPs]; CCSSI, 2010). The SMCs describe mathematics content at each grade level that students ought to know; whereas the SMPs characterize mathematical behaviors and habits that students across K-12 grade levels ought to demonstrate during instruction (CCSSI, 2010). Problem solving is found throughout every grade level in SMCs and also as the first SMP. Thus, there is a strong emphasis on problem-solving. The PSMs are intended to be measures that assess students' mathematical problem-solving performance in ways that are connected to grade-level SMCs and SMPs. Readers interested in greater details about PSMs development, connection to Standards (CCSSI, 2010), and the validation process associated with developing the PSMs should consult prior work (e.g., Bostic and Sondergeld, 2015, Bostic et al. 2017, 2018, 2019a, 2020). Items on the PSMs are linked with mathematics content found in the SMCs and SMPs. There are 12 grade-level items on the PSM3, PSM4, PSM5, and PSM6 while PSM7 and PSM8 each possess 15 items. Data for the present study focus on PSM3. An example item from the PSM3 reads: "Beth is coloring a picture using crayons. The box of crayons has 6 blue crayons, 4 yellow crayons, 8 green crayons, and 6 red crayons. What fraction of the box of crayons is green?" The 1–1 think alouds and WCTAs conducted for the present study are situated within the scope of a broader research project on developing and validating PSMs. Prior 1–1 think alouds had been conducted nearly 2 years before the present study with an aim to gather response process evidence. Those prior think alouds with different students helped to inform data selection and procedures for the WCTAs. All names used in the manuscript are pseudonyms.

3.2. Participants and Setting

Multiple districts had been representatively and purposefully selected to participate in test development because of their student population (diversity of racial and ethnic identities) and geographic location (rural, suburban, and urban). All work for the WCTA study took place in one suburban school district with a low student poverty rate and average student population size when compared to other districts in the Midwest state located in the USA (Ohio Department of Education, 2013). Our team sought to try WCTAs in one setting before moving to additional districts. Additionally, a strong, positive relationship with the administrators and teachers in this chosen district guaranteed that WCTAs might happen as planned. Third-grade students and the PSM3 data were selected for the present study because this research team intended to explore WCTAs as a means to gather data from elementary school-aged students and none of the students had been previously exposed to the tasks from a prior year unlike the other grade levels.

There are two sets of third-grade participants involved in the study: one set for 1–1 think alouds and another for WCTAs. The first set of student participants for 1–1 think alouds consisted of 12 students. At least two students representing each ability level participated in 1–1 think alouds. The authors sought below-average, average-, and above-average ability students for both groups. Mathematics ability level was associated with students' grade-level ability in mathematics. These ability-level judgments were made by classroom teachers and based upon class grades, prior assessment scores, and progress monitoring data. The sample for 1–1 think alouds was reasonably and representatively sampled in terms of gender, mathematics ability level, as well as racial and ethnic identity. Lastly, these 1–1 think alouds were conducted in quiet settings such as a school meeting space or library.

One year following the 1–1 think alouds, researchers returned to the same district to conduct WCTAs. Third-grade students that participated in the 1–1 think alouds previously matriculated and became fourth-grade students during WCTA data collection, so there was no chance that students might see the same problem. Third-grade classes had 16 students each and participation across the school included two grade 3 classes.

3.3. Procedures

3.3.1. Prior Work

Item characteristics were taken from content expert's input as well as item characteristics drawn from analysis of 72 participants' data during a pilot study. Content experts included grade-level mathematics teachers with three or more years experience in that grade level, university-level mathematics educators whose expertise included elementary mathematics, and university-level mathematicians who were familiar with the Standards (CCSSI, 2010). For this study, mathematics educators describes individuals with a PhD or EdD, who teach education courses to preservice and inservice mathematics teachers at a four-year institution. Mathematicians describe individuals who held a PhD in mathematics and teaching responsibilities were in a mathematics department. These experts provided individual reports that were analyzed using thematic analysis (Creswell, 2011; Hatch, 2002), and are shared in prior research (see Author, Bostic at al. 2018). Pilot study data were scored dichotomously and analyzed using Rasch modeling (1960/1980). Those results from pilot study data supported creating tertiles (i.e., less difficult items [$-3 \leq$ item difficulty ≤ -1], average difficulty items [$- 1 <$ item difficulty ≤ 1], and more difficult items [$1 <$ item difficulty ≤ 3]). A review of qualitative and quantitative findings led to agreement on item characteristics to use in the think alouds, that is: less difficult, average difficulty, and more difficult items. Readability had been previously examined (Bostic et al. 2018) and all items had appropriate grade-level readability scores; moreover, teachers confirmed that the items were developmentally appropriate for third-grade students.

3.3.2. One-on-One (Traditional) Think Aloud

Prior to the 1–1 think alouds, researchers created packets of four problems to use during the 1–1 think aloud. Additionally, teachers shared ability-level perceptions of their students with the researchers. Two researchers conducted the 1–1 think alouds. One researcher met the student in the classroom and escorted the child to a quiet room. Teachers, administrators, and researchers agreed on 20-min intervals for students to work on three or four problems because it took 5 min for students to walk to the 1–1 think aloud meeting room and another 5 min to return to class for a total of 30 minutes out of class. Each 1–1 think aloud took as few as 12 minutes or as many as 35 minutes, with the average think aloud duration totaling approximately 20 minutes. A video camera was situated so that it captured students' work on the problem and their voice, but not their faces. The researcher sat across from the respondent and made notes during the 1–1 think aloud.

A researcher explained the directions of the 1–1 think aloud to the respondent. Students were asked to share their thoughts by thinking aloud and write ideas on the paper. The researchers conducting 1–1 think alouds expressed that they could not answer questions. Students were allowed to spend a maximum of 8 minutes on any item, which was derived from content expert feedback and think alouds conducted earlier in the same grade level. The researcher encouraged respondents to show all of their work and not to erase ideas unless it was necessary. Each participant read the problem aloud. Participants were reminded to verbally share ideas when they appeared pensive but were not saying anything. Each problem was printed on a single piece of paper. Students were given a practice problem, then after completing it, asked if they had questions about the process. The respondent returned the practice problem to the researcher, who in turn gave the respondent the first problem. This continued until all problems in the problem set were completed. At the end of the 1–1 think aloud, the researcher asked the respondent: “Would you like to share anything about the think-aloud process with me?” Students were escorted back to their classroom following their response to this question.

3.4. Whole-Class Think Aloud (WCTA)

WCTAs in this study were conducted in classrooms during normal classroom instruction. Two researchers and one classroom teacher were present in the classroom during implementation. Students from the school in the same district as those that participated in 1–1 think alouds, participated in WCTAs. Classroom teachers were told the general content of the problems (e.g., operations with fractions, geometry, and data/measurement) just prior to WCTA administration but the problems themselves were not seen until the WCTA began. WCTAs in this study started with the classroom teacher introducing the class to the two researchers in the room. The same researchers who carried out the 1–1 think alouds also carried out the WCTAs, in order to maintain similarity across data collection. Teachers told students that they would work on word problems for about 30 minutes, which is their normal period of time for math lessons. They had 6 min to solve each item; that amount of time was based on prior 1–1 think alouds. If students finished early, then they could work on a second approach to solve each item. If every student finished early, then the class return the task to the researchers and moved to the next task.

Teachers explained that thinking aloud meant any ideas in one’s mind that seemed relevant to the problem ought to be voiced aloud. Students were encouraged to write their ideas on paper and to think aloud with a partner. Teachers also explained that they would actively listen during the WCTA and not provide feedback, hints, or answer questions during it. Language from one teacher, named Emily, was typical of her response: “I want to hear what you think and see your work on the problems.” Next, the teacher grouped students based upon similar grade-level mathematics ability levels (e.g., below-average, average, and above-average ability) as described earlier. Teachers were instructed to use their best judgment of students’ mathematical abilities based upon grades, prior assessment scores, and progress monitoring data. Due to the number of students in each ability category, it was not always possible for teachers to have same ability groups in all cases. However, in every WCTA experience, at least 85% of the students were grouped with similar-ability peers. When there was mixed ability pairing, the teacher made certain above-average and below-average students were not placed together because of a fear that the above-average student might dominate the problem-solving conversations. The classroom teacher shared students’ ability levels with the researchers. In every classroom, there were male-male, female-female, and male-female groupings.

Each problem was printed on a single piece of paper so that each student received a copy of it. Then, students received a practice problem. Each student received the same problem. The teacher reminded students that they should talk aloud to their partner while problem-solving. Students could write or draw on their papers. Researchers stressed that it was important students showed all their work, even if they changed strategies while problem-solving. They were to cross out work they did not intend to finish rather than erase all of it. The final direction came from the teacher: a reminder that the teacher and students would chorally read the task before starting. This directive served as a means to confirm that students could read the task, similar to how students read each task in the 1–1 setting.

Following the choral reading, students began working on the task. The two researchers and teacher circulated throughout the room. Researchers took notes about students’ problem solving, words spoken aloud, and made notes about what various ability leveled groups of students said. The teacher did not give hints, help, or strategies to students. Researchers and the teacher observed and took notes. Teachers shared that their notes were intended for reflection about possible instructional revisions. Similar to the 1–1 think aloud procedures, a researcher asked the students one question at the conclusion of the WCTA: “Would you like to share anything about the think-aloud process with me?”

3.5. Data Collection

All students that participated in general education (i.e., push-in) mathematics instruction were eligible to participate in the 1–1 instruction. Data from students who received accommodations or modifications to instruction or assessments were excluded from analysis.

This 1–1 think aloud process allowed for four problems to be administered and as many as 6 problems during the WCTAs. Data for this study focused on two of three tertiles from PSM3 data – less difficult and average difficulty items. There are reasons the authors chose to focus on responses to those items. First, data were gathered in both 1–1 and WCTA settings in January, which is approximately the middle of the academic year. Second, we consulted with grade-level teachers prior to think alouds; they insisted that content found in the more difficult items was examined later in the academic year. Their responses were grounded in their judgment of students' current mathematical problem solving as well as results from progress monitoring tools that connected to the Standards (CCSSI, 2010). Thus, administering more difficult items to students might likely lead to a low response rate. Finally, we drew upon past think-aloud data collected from the end of an academic year. We were confident that students were not prepared to solve the more difficult items during the middle of the academic year.

3.6. Data Analysis

There were two parts for the data analysis. The first part was analyzing 1–1 think aloud and WCTA data separately to investigate correctness and strategy use. All think-aloud problems were scored dichotomously (correct/incorrect). Dichotomous scoring has been used with prior PSM research (e.g., Bostic and Sondergeld, 2015, Bostic et al. 2017) and recent research demonstrated that there was no difference between partial-credit and dichotomous scoring (Sondergeld et al., 2020). A sample of 10 problems were scored by two researchers using an answer key. Their results had complete scorer agreement, which suggested the scoring process met expectations.

A second procedure was used to investigate the viability of students' strategy use. Viability of a strategy, for this study, was defined as (a) drawing upon appropriate mathematical representations and (b) using mathematical procedures that may lead to a solution. A viable strategy may include arithmetical errors. Viability of a strategy was coded first by problem solvers' representation uses. Mathematical procedure coding was coded based upon known ways to reach a solution garnered from expert feedback provided by third-grade teachers, university-level mathematics education faculty who previously taught in the elementary grade levels, and a mathematician with education experience. All three groups of experts provided descriptions of procedures for each item that were mathematically sound and developmentally appropriate.

Past research on students' problem-solving performance (see Yee and Bostic, 2014) indicated that more successful elementary-aged problem solvers tend to use multiple representations (e.g., symbolic and pictorial representations) while solving mathematics word problems. Hence, that work informed the representation coding protocol for the present study. Students' representation use on the items were coded as symbolic, pictorial, tabular, verbal, or a combination of these approaches called mixed (see Table 2). These representations were selected because Lesh and Doerr (2003) describe this categorization as appropriate ways to categorize mathematical representations.

It also provided further evidence for comparison across the 1–1 think aloud and WCTA approaches. Two researchers independently coded 10 problems using the protocol then compared results. The coding process started with Table 2 and seeking exemplars for each representation code. These exemplars provided ideas about how the codes might be reified in the PSM data. The second step was to practice coding with a random sample. Ten work samples from different items were randomly pulled from the larger data set. Each researchers worked independently then compared codes. There was complete agreement across coders related to problem solvers' representation uses. Representation coding continued and any concerns were brought to

Table 2. Representation coding protocol.

Representation Code	Representation
A	Symbolic: Use of abstract symbols (i.e., variables) or number expressions
B	Pictorial: Use of pictures or diagrams
C	Tabular: Use of tables or charts
D	Verbal: Use of written words
E	Mixed: Symbolic & Pictorial
F1	Mixed: Symbolic & Tabular
F2	Mixed: Symbolic & Verbal
F3	Mixed: Pictorial & Tabular
F4	Mixed: Pictorial & Verbal
F5	Mixed: Tabular & Verbal

the other coder for discussion. A similar approach was used to code the procedures students used. Again, two researchers coded students' procedures using the guides provided from the expert panel. A small sample of 10 problems was taken to check for agreement across coders and results indicated complete agreement. There was complete agreement across the two coders related to procedure use. Results from the performance and strategy analyses were intended to highlight information about the viability of students' problem solving on the items used during the 1–1 think alouds and WCTAs. These results are presented related to performance on the problems and viability of the mathematical strategies used in solving them.

The second part of data analysis was comparing 1–1 think aloud and WCTA qualitative data. All work was videotaped and later transcribed, when audible. Visual data such as videos and audio data such as respondents' words were analyzed using inductive analysis (Creswell, 2011; Hatch, 2002). WCTA data and 1–1 think-aloud data were analyzed separately and later compared to understand the similarities and differences in results from the analytic approach. An aim for inductive analysis is to inductively draw out themes or key ideas based upon available data (Creswell, 2011). The team used seven steps for inductive analysis. First, researchers became familiar with the available data for analysis. Researchers reviewed WCTAs and 1–1 think aloud data as well as students' problem-solving. Step two was to rewatch videos to clarify any ambiguity that arose during the first review of data. The third step was making notes about students' problem solving based upon the available data. Step four aimed to categorize notes into general ideas. Fifth, discussions about categories that might be eliminated or revised based upon the findings occurred. Step six was to review each category and consider the amount and quality of evidence related to it. Those categories with two or more pieces of counterevidence or a paucity of evidence were deleted from the analysis. The final seventh step involved drawing those categories where possible into a broad theme or general idea. Ideas that remained as themes were retained for comparison across WCTA and 1–1 think alouds.

4. Findings

A first aim is to describe results from the 1–1 think alouds and WCTAs separately. The second aim is to contrast the two approaches and share themes. Results from both WCTA and 1–1 think alouds were similar, providing evidence that students were responding in anticipated ways. Findings from the 1–1 think alouds and WCTAs were compared to explore differences in the content shared or the data collection process. Broadly speaking, there were three key themes. The first theme addresses the research question: How do findings from 1–1 and WCTAs compare? Results from both types of think alouds led to similar student responses across the 1–1 and WCTAs in this study. The remaining themes address the research question: How do 1–1 and WCTAs compare as data analytic tools for gathering response process data? A second theme was that those who participated in WCTAs shared qualitatively richer responses than traditional 1–1 think-aloud respondents both in terms of oral and written communication. A third theme was that students expressed feeling more comfort during the WCTAs than in the traditional 1–1 think aloud setting.

Table 3. Problem-solving performance and viability of strategy (%).

	1–1 Think Aloud		WCTA	
Item Difficulty	Correct Solution	Viable Strategy	Correct Solution	Viable Strategy
Average Difficulty	8%[a]	50%[a]	10%[b]	63%[b]
Less Difficult	34%[a]	59%[a]	44%[b]	57%[b]

[a]sample size = 12
[b]sample size = 32

4.1. Theme #1: Similar Findings across 1-1 Think Alouds and WCTAs

Table 3 communicates results related to anticipated problem-solving performance and strategy viability. There was reasonable evidence drawn from the results suggesting similar response patterns between 1–1 think alouds and WCTAs.

On average, 8% of students in 1–1 think alouds correctly responded to items of average difficulty compared to 10% of respondents in WCTAs. Similarly, 34% of students in 1–1 think alouds correctly answered less difficult items compared to 44% of students in WCTAs. There was a similar pattern when examining viability of problem-solving strategies. Fifty percent of students in 1–1 think alouds provided viable strategies for average-difficulty items compared to 63% of students in WCTAs. Examining the attempts on less difficult items, 59% of students in 1–1 think aloud settings offered viable strategies compared to 57% of students in WCTA settings. Chi-square analyses were performed on the correct solution and viable strategy 1–1 think aloud and WCTA data. Chi-square results were not statistically significant ($p > .05$), indicating that there were no significant differences between 1–1 think aloud and WCTA respondents' outcomes. Taken collectively, these findings suggest that students in both think-aloud settings had similar problem-solving performance and strategy use.

4.2. Finding #2: Greater Sharing during Think Alouds

During both 1–1 and WCTAs, students were reminded to share their thinking aloud while problem-solving. Students in the 1–1 think aloud setting tended to need three or more reminders during the 1–1 think aloud, averaging to one reminder during each task. Respondents usually quipped back "Oh yeah, sorry I forgot" or "Oh ok" after being reminded. This was typical across respondents of varying gender, racial, and ethnic identity, as well as ability level. On the other hand, respondents in the WCTA setting were rarely reminded to think aloud after the practice task. Videodata from the WCTAs indicated that a researcher typically reminded one pair of respondents to think aloud during the entire WCTA. A researcher typically reminded a respondent during the 1–1 think alouds to share thinking aloud more often than in the WCTA setting.

There was consistent talk about problem-solving during the WCTAs. When a group of WCTA participants communicated that they had solved the problem and there was remaining time for others to keep working, one of the researchers or a teacher encouraged students to attempt solving it using a different approach. Students in the 1–1 setting who responded to the task in fewer than 5 min were also encouraged to explore a different problem-solving approach for that task. Generally speaking, students from the 1–1 think aloud setting who completed the task were reticent to explore a second problem-solving strategy and tended to ask if they could move to the next task. They asked questions like "Can I just go to the next problem" or said statements such as "I solved the problem and don't want to keep going." This contrasted with students in the WCTA setting who were much more willing to investigate a second problem-solving strategy. While second problem-solving strategies were not usually completed in the allotted time in either think aloud setting, WCTA respondents had different comments when asked to explore the problem further. Respondents in WCTA settings commented after learning that they were interested to seek another strategy, "Alright, I bet we [talking to the other respondent] can figure out another way to do this [problem]" and "I like solving problems like this.

Let's try another way and see if we get the same answer as before." There was palpable excitement from respondents in the WCTA setting compared to the 1–1 think aloud setting and altogether, a willingness to externalize problem-solving during the WCTA that was not as present in the 1–1 think aloud setting.

4.3. *Finding #3: Comfort in Think-aloud Setting*

There was a noticeable difference in students' comfort level during the think aloud. Specifically, students in the WCTA setting expressed feeling more comfortable thinking aloud during the WCTA compared to the 1–1 think aloud setting. Students in the 1–1 think aloud setting expressed feeling some anxiety during the beginning, and the approach felt unnatural. Audrey shared at the end of her think aloud that "I don't know you [researcher] so it's a little awkward at first. . . . it's OK now [end of think aloud] but still weird." Marcus said something similar at the end of his 1–1 think aloud: "I talk to people in my class while I'm doing math problems . . . but it's weird that you want me to talk to you [researcher]. I just met you a few minutes ago." Finally, Elliot said, "I don't want to be rude but I don't know you. It's a little strange, you know, sitting at a table and I don't know you. I feel weird thinking aloud because I'm not talking to anyone. You [researcher] can't say anything." These comments were typical from the 1–1 think aloud participants – they expressed feeling that the 1–1 experience was unnatural and somewhat awkward, which differed from WCTA findings.

Students in WCTAs did not communicate such anxiety and actually were excited to think aloud with peers about mathematics problems. At the end of one WCTA setting, Nikolas told a researcher that "This was fun. I can't believe we did all those problems in 30 minutes! I like doing hard problems and being able to tell someone what I'm thinking." Margret shared that she wanted to do more WCTAs: "Wait, we're all done? This is fun. . . . I like being able to do math while talking about it with another person. Do we get to do more [WCTAs] tomorrow?" Parallel to these comments, Hattie told a researcher that "This [WCTAs] feels like the normal stuff we do in class." Taken collectively, these comments from WCTA participants indicated that they felt comfortable thinking aloud during the WCTA setting and furthermore, WCTA respondents felt more comfortable in that setting than those in the 1–1 think aloud setting.

5. Discussion

The first theme addresses the research question: How do findings from 1–1 and WCTAs compare? Results from both types of think alouds led to similar student responses across the 1–1 and WCTAs in this study. The remaining themes address the research question: How do 1–1 and WCTAs compare as data analytic tools for gathering response process data?

WCTAs are framed as a data analytic tool meant to capture evidence about a respondent's thinking. WCTAs may foster comparisons to intended ways of thinking that were expected by an assessment development team or content expert panel. We offer three reasons to consider WCTAs and take up how they might connect to both theory and practice related to validation studies exploring response process evidence. Those three reasons are: (a) developing a new data collection tool for gathering response process evidence; (b) changing the think-aloud environment; and (c) fostering partnership development between assessment developers and schools. An implication from this study is that WCTAs may serve as tools for future data collection focusing on mathematics items. A second implication is that the WCTAs positioned the classroom teacher as a collaborator in the data collection process. Teachers involved in the WCTAs shared ways they intended to change instruction as a result of what they heard and saw. Limitations and delimitations for the present study will be shared, which may foster future researchers to further investigate WCTAs.

5.1. Developing a New Data Collection Tool for Response Processes

Think alouds in a 1–1 setting have been routinely described as a viable means to gather response process data (AERA et al., 2014; Fonteyn et al., 1993; Leighton, 2017). Other tools to gather response process data include eye tracking and fMRI technologies (AERA et al., 2014; Schindler & Lilienthal, 2019). The present study provides a new tool in the arsenal for those seeking to gather response process data. WCTAs were used to better understand respondents' cognitive activity while engaged in problem-solving using developmentally appropriate mathematics problems. We contend that WCTAs be used to augment previously collected 1–1 think aloud data or to bring a different perspective on the cognitive processes respondents use while working on mathematics problems. Results from WCTAs may serve to support validity arguments instrument developers create.

Drawing from the first finding, there were not substantial qualitative differences in performance or strategy use across think-aloud settings. There were increases in performance in outcomes from the WCTA setting on less difficult items compared to 1–1 think aloud settings; albeit, the differences were not statistically significant. This suggests that respondents performed similarly across the two settings. We hold the findings somewhat tentative because of the nature of the small samples used in this study. It is plausible that students in the WCTA setting may have higher performance because they were talking to one another, which may increase the likelihood for correct responses and viable strategies. Using the 1–1 think aloud findings as the benchmark for comparison, WCTA results indicated that students performed statistically similar on items compared to respondents of similar ability in the 1–1 think aloud setting. At the same time, the percentage of viable strategies across think-aloud settings was similar. While there were slightly more viable strategies displayed on average difficulty items during the WCTA than 1–1 think aloud setting, slightly more respondents in the 1–1 think aloud setting used viable strategies on less difficult items than respondents in the WCTA setting. In both cases, approximately half of the respondents in each think-aloud setting used developmentally appropriate strategies. Taken collectively, these quantitative findings tend to suggest that the WCTA approach draws out similar information about respondents' mathematical problem-solving performance and strategy use as the 1–1 think aloud approach.

5.2. Think-aloud Environment and Richness of Data

As described in the literature, recommended settings and contexts of think alouds are sterile, quiet conditions whereby the respondent and researcher are able to interact without distractions (Fonteyn et al., 1993; Leighton, 2017; Wilson & Miller, 2014). The intent for sterile settings is that such a setting free from distractions offers the researcher a better perspective on the respondent's cognitive processes during the think aloud. It is easy to hear ideas and see the work at the same time when in a quiet space. A challenge with this approach is that a different environment from the classroom may interfere with a respondent's thinking because it is novel from a normal testing environment where tests are usually administered. For instance, the PSMs are designed to be administered in classrooms during normal school settings. As a result, data from the WCTA setting seemed richer in language and provided ideas about students' mathematical problem solving that were not present during the 1–1 think alouds.

A finding from the present study was that respondents felt more comfortable in the WCTA setting and needed fewer reminders to share their mathematical thinking during 1–1 think-alouds. Test developers aiming to create mathematical problem-solving assessments used in classroom settings might consider using WCTAs in conjunction with other techniques to understand cognitive processes. The WCTA approach offered a naturalistic setting that was more comfortable – according to respondents – than the 1–1 think aloud setting. It is viable that the WCTA offers a natural and comfortable setting such that respondents are better able to convey their responses than in a 1–1 setting. In the context of the present study in a classroom setting, third-grade students' knowledge may be situated within the community; the idea that knowledge and learning, including mathematics work, is situated or part of a shared experience. This view on learning and knowledge as being situated within the classroom is not new (e.g., Cobb & Bowers, 1999; Lave & Wenger, 1991). Thus, WCTAs may connect with these theories of learning and knowledge.

As a result of this work with WCTAs, respondents seemed more willing to share their thinking without prodding from a researcher administering the items. For this reason, it is worth exploring in future research, whether or not this difference in the comfort students felt and how well they shared ideas was influenced by other factors. In both think-aloud settings, the researchers were new to the students; however, the teacher was present in the room during the WCTA setting. With the teacher present in the room, the teacher may choose to leverage the response process evidence for their future instruction if the assessment items connect to classroom learning objectives.

5.3. Partnerships with Schools

Developing assessments and tests for K-12 classroom use usually involves working with potential respondents, who are K-12 students. Access to K-12 students usually begins by interacting with administrative staff such as superintendents, curriculum coordinators, and/or principals. Then, access to K-12 students is mediated through scheduling with classroom teachers and administrative staff. Gathering response process data, like in a 1–1 think-aloud setting, may involve pulling a student away from everyday classroom instruction. Thus, validation work requires developing strong partnerships with school districts, including administrators and teachers, in order to build the trust that the benefits of engagement in the think aloud outweigh the challenge of missing everyday classroom instruction.

The research team did not share the items used in the WCTA with the mathematics teachers until the day of the WCTA administration. This limited the chance that students might be inadvertently exposed to the items, hence retaining item security. During the study, classroom teachers frequently commented that they wished they could observe the think alouds. While we offered to watch the videos of 1–1 think alouds of students to their teachers, teachers expressed that they did not have time to watch multiple 30-min segments. On the other hand, they were excited to partner in the WCTA experience. One teacher told the research team before the WCTA experience that “I want to have the opportunity to listen and watch students’ thinking. Since the problems are connected to content, I don’t feel I’m losing an instructional day with any of them [students].” After the WCTA experience, the same teacher shared “I use formative assessment but I’ve never done anything like this. It [WCTAs] gave me information about what my students think and what they remember from last year. I will absolutely be making changes to my instruction based upon this [WCTA].” As a result of the WCTA experience in third grade, the research team was invited to lead a professional development meeting focused on WCTAs to support teachers in other grade levels. The third-grade teachers involved in the present study wanted to conduct their own WCTAs as part of their instruction. The WCTA experience provided a means to enhance the school–university relationship forged in part of the broader research agenda aimed at developing new problem-solving measures. A result of the present WCTA study focused on scholarly work, was conducting a parallel study on practical implications for teachers (see Hicks and Bostic, 2020). One practical implication from conducting WCTAs was that teachers felt connected with the assessment development process and garnered a new way to gather data about students’ cognitive processing that they had not accessed previously through other formative assessment approaches. Ultimately, WCTAs helped to strengthen the school–university partnership and maintain objectives for the larger project focused on assessment development of a series of mathematics measures.

Building strong relationships with teachers, administrators, and schools are important to test developers aiming to make validity arguments that describe score interpretations for proposed uses (Kane, 2012, 2013, 2016) related to in-class tests that explain how the results and interpretations should be used. For the purposes of this study, having a strong school–university relationship allowed the research team to trust that the ability-level groupings put together by classroom teachers were appropriate. A strong school–university partnership also has potential to spur partnerships including future test development and modifications as well as teachers in one district sharing with others about the positive work with those associated with test development. In sum, the WCTA experience around the mathematical problem-solving problems was viewed as positive by students, teachers, and administrative staff.

5.4. Delimitations, Limitations, and Future Research

Because an aim of this study is to examine a new methodological approach, we want to clarify ways that the study was delimited and there are some limitations with the findings themselves. First, this study and its findings are intentionally delimited to work with elementary-aged students and mathematical problem-solving items. We limit our generalizations from this study to mathematics problems with elementary students – future study is warranted to explore WCTAs with other tasks. While we hypothesize that success with third-grade students suggests that success in later grade levels is likely, such research must be fully examined. Second, future work might be done with multiple school districts, school districts representing greater racial or ethnic diversity, and more grade levels within the context of WCTAs. This design and development study was conducted using the Common Guidelines for Education and Research Development (Institute of Education Sciences, U.S. Department of Education, and National Science Foundation, 2013) to generate a viable alternative or supplement to 1–1 think alouds. Further study under efficacy trials are warranted to better understand the appropriate uses and limitations of WCTAs. Results from such a study comparing 1–1 and WCTA outcomes of students from different schools might provide greater saturation for better understanding how the WCTA functions in different learning contexts and different tasks. A third future study ought to explore WCTAs with respondents of different ages (e.g., secondary students or adults) as well as different content (e.g., biology, history, or English). A case study might investigate how teachers that participated in WCTAs might use them as part of their formative assessment routines, what they learn from the WCTAs, and how they use those data to make instructional decisions.

This study was intentionally delimited to include developmentally appropriate mathematical problem-solving tasks. Each constructed-response task had a single correct response that might be reached through at least three different developmentally appropriate strategies, as communicated by the expert panel. Outcomes from WCTAs might differ if respondents are expected to select and justify a choice from a multiple-choice task. Similarly, findings might be different if there were multiple correct responses for a given task.

An important limitation with the WCTA approach is that it was not feasible to listen to conversations from 10 to 13 pairs of students in the classroom at one time using available technology. There were three adults in the room, two of which were taking notes. Videorecorders were placed on two sides of the classroom to capture as much information as possible for review. Researchers strategically positioned themselves to listen to differing ability students during the WCTAs, and shifted to listen to different groups of students for each problem. A future research study might investigate a WCTA setting such that videorecorders are placed near each pair of students, and findings from those WCTAs be compared to respondents outcomes from 1–1 think-aloud settings.

Acknowledgments

Ideas in this manuscript stem from grant-funded research by the National Science Foundation (NSF 1644214, 1720646,1720661,1920619,1920621). Any opinions, findings, conclusions, or recommendations expressed by the authors do not necessarily reflect the views of the National Science Foundation. The authors would like to acknowledge individual_A and individual_B for their support in conducting 1-1 think alouds prior to this study. Additionally, the authors are grateful to the *Applied Measurement in Education* editorial team for their feedback on this manuscript as well as individual_C for her encouragement related to this manuscript.

Funding

This work was supported by the Directorate for Education and Human Resources [1644214,1720646,1720661, 1920619,1920621].

ORCID

Jonathan David Bostic http://orcid.org/0000-0003-2506-0491
Toni A. Sondergeld http://orcid.org/0000-0001-7264-5607

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Bostic, J.,Matney, G., Sondergeld, T., & Stone, G. (2018, November). *Content validity evidence for new problem-solving measures (PSM3, PSM4, and PSM5)*. In T. Hodges, G. Roy, & A. Tyminski (Eds.), *Proceedings for the 40^{h} Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 1641). Greenville, SC.

Bostic, J., Matney, G., Sondergeld, T., & Stone, G. (2019a). *Developing a problem-solving measure for grade 4*. In M. Graven, H. Venkat, A. Essien, & P. Vale (Eds.), Proceedings of the 43rd Meeting of the International Group for the Psychology of Mathematics Education (Vol. 4, pp. 4–16). Pretoria, South Africa.

Bostic, J., Krupa, E., & Shih, J. (2019b). Introduction: Aims and scope for Assessment in mathematics education contexts: Theoretical frameworks and new directions. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 1–11). New York, NY: Routledge

Bostic, J., Matney, G., Sondergeld, T., & Stone, G. (2020). Measuring what we intend: A validation argument for the grade 5 problem-solving measure (PSM5). Validation: A Burgeoning Methodology for Mathematics Education Scholarship. In J. Cribbs & H. Marchionda (Eds.), *Proceedings of the 47th Annual Meeting of the Research Council on Mathematics Learning* (pp. 59–66). Las Vegas, NV

Bostic, J., & Sondergeld, T. (2015). Measuring sixth-grade students' problem solving: Validating an instrument addressing the mathematics common core. *School Science and Mathematics Journal*, *115*, 281–291. doi:10.1111/ssm.12130

Bostic, J., Sondergeld, T., Folger, T. & Kruse, L. (2017). PSM7 and PSM8: Validating two problem-solving measures. *Journal of Applied Measurement*, *18*(2), 151–162.

Cobb, P., & Bowers, J. (1999). Cognitive and situated learning perspectives in theory and practice. *Educational Researcher*, *28*(4), 4–15.

Common Core State Standards Initiative. (2010). *Common core standards for mathematics*. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Creswell, J. (2011). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson.

Fonteyn, M., Kuipers, B., & Grobe, S. (1993). A description of think aloud method and protocol analysis. *Qualitative Health Research*, *3*(4), 430–441. doi:10.1177/104973239300300403

Hatch, J. A. (2002). *Doing qualitative research in educational settings*. Albany: State University of New York Press.

Hicks, T. & Bostic, J. (2020, March). *Assessing for misconceptions using whole-class think alouds*. Poster presented at annual meeting of the Research Council on Mathematics Learning. Las Vegas, NV.

Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In B. Zumbo & A. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 1–12). Cham, Switzerland: Springer.

Institute of Education Sciences, U.S. Department of Educartion, & National Science Foundation. (2013, August). *Common guidelines for education research and development*. Retrieved from https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf13126

Kane, M. T. (2012). All validity is construct validity. Or is it? *Measurement: Interdisciplinary Research and Perspectives*, *10*(1–2), 66–70.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. doi:10.1111/jedm.12000

Kane, M. T. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (Vol. 2, pp. 64–80). New York, NY: Routledge.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*(1), 127–135.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cabridge, UK: Cambridge University Press.

Lavery, M., Jong, C., Krupa, E., & Bostic, J. (2019). Developing an assessment with validity in mind. In J. Bostic, E. Krupa, & J. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 12–39). New York, NY: Routledge

Lavery, M. R., Bostic, J. D., Kruse, L., Krupa, E. E., & Carney, M. B. (2020). Argumentation Surrounding Argument-Based Validation: A Systematic Review of Validation Methodology in Peer-Reviewed Articles. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12378

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*(4), 6–15. doi:10.1111/j.1745-3992.2004.tb00164.x

Leighton, J. P. (2017). *Using think aloud interviews and cognitive labs in educational research*. Oxford, UK: Oxford University Press.

Lesh, R., & Doerr, H. (2003). Foundations of a models and modeling perspective on mathematics teaching, learning, and problem solving. In R. Lesh & H. Doerr (Eds.), *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching* (pp. 3–33). Mahwah, NJ: Erlbaum.

Ohio Department of Education. (2013). *Typology of Ohio school districts*. Retrieved from https://bit.ly/2Bbvdwi

Padilla, J.-L., & Benitez, I. (2014). Validity evidence based on response processes. *Psichotherma*, *26*, 136–144.

Plake, B. S., & Wise, L. L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing? *Educational Measurement: Issues and Practice*, *33*(4), 4–12. doi:10.1111/emip.12045

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press.

Schindler, M., & Lilienthal, A. J. (2019). Domain-specific interpretation of eye tracking data: Towards a refined use of the eye-mind hypothesis for the field of geometry. *Educational Studies in Mathematics*, *101*(1), 123–139. doi:10.1007/s10649-019-9878-z

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicotherma*, *26*(1), 100–107.

Sondergeld, T., Stone, G., Kruse, L., Bostic, J., & Matney, G. (2020). *Evaluating Dichotomous and Partial-Credit Scoring within a Constructed-Response Assessment: Is More Information Always Psychometrically Better?* Paper presented at annual meeting of the annual meeting of the American Education Research Association. San Francisco, CA.

Wilson, S., & Miller, K. (2014). Data collection. In K. Miller, S. Wilson, V. Chepp, & J.-L. Padilla (Eds.), *Cognitive interviewing methodology* (pp. 15–34). New York, NY: Wiley.

Yee, S., & Bostic, J. (2014). Developing a contextualization of students' mathematical problem solving. *Journal of mathematical Behavior*, *36*, 1–19. doi:10.1016/j.jmathb.2014.08.002