Abstract

A think aloud lab is a method used to gather evidence based on response processes to determine whether tests have construct validity. The current study sought to gather evidence based on response processes during a think aloud lab which followed a protocol that Johnstone, Bottsford-Miller, and Thompson (2006) developed. The data helped to evaluate whether TNReady science test items evoked the intended cognitive and problem-solving processes, barriers that may have impeded their access (e.g., high reading loads), and whether or not valid inferences could be drawn from the results in regard to students’ knowledge of science content. A total of 33 students across three grades (fifth, eighth, and high school) participated in the think-aloud lab. Students were asked to talk out loud about their thinking as they completed science items. Students were also asked to complete Maze reading probes and a five-item survey after the test. Survey items focused on student opinions in regard to difficulty and usability. Results indicated that the TNReady science test items evoked the intended cognitive processes. Participants interacted with the tests as the developers intended. The cognitive lab procedures are outlined, and key findings are highlighted. Theoretical influences, such as accessibility theory, cognitive load theory, and universal design standards, as well as previous think aloud studies, are reviewed. Limitations of the current study and implications for practice are also discussed.
Acknowledgements

I would like to thank my dissertation chair, Dr. Ryan Kettler, and my committee member, Dr. Kelly Feeney-Kettler. I am supremely grateful for the opportunities to join your research lab and travel to Nashville for this project. Thank you both for your support and feedback throughout this process. Dr. Kettler, your mentorship has taught me so many valuable lessons, including how to be more confident in myself and to take risks. Thank you for challenging me to push myself in scholarly pursuits and in life. Dr. Feeney-Kettler, I have so much admiration for your dedication to your career and your family. Thank you for being a superb example of balancing professionalism and motherhood.

To all of my supervisors along the way, thank you for everything you have shown and taught me. You all have helped to shape my development as a professional and as a human being these last four years, and for that I am extremely grateful.

To my husband, Michael, thank you for your unending encouragement and faith. I would not have taken this leap or have gotten this far without you. You are my anchor, and the light and love of my life. To my parents, thank you for instilling your values of education and excellence, and for your continuous love and support. Thank you for keeping me grounded and reminding me of my roots while pushing me to fly. To my sister and brother-in-law, Kelliann and Frank, thank you for your support and understanding throughout this journey. Kelli, thank you so much for your expertise and assistance. To Emily, David, Stephanie, Leon, the rest of my family, and friends – thank you for the family dinners, texts and emails, and cheering me on along the way.

To my son, Luca Arthur, thank you for choosing me to be your mother. Thank you for bringing so much joy and teaching me the most important life lessons. You are my greatest gift and blessing, and I loved you before I met you.
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-Test Survey</td>
<td>48</td>
</tr>
<tr>
<td>Limitations</td>
<td>49</td>
</tr>
<tr>
<td>Implications for Practice</td>
<td>49</td>
</tr>
<tr>
<td>Conclusions</td>
<td>50</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>52</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Specifications of Think Aloud Studies of Achievement Tests</td>
<td>16</td>
</tr>
<tr>
<td>2. Demographics of the Sample by Grade Band</td>
<td>26</td>
</tr>
<tr>
<td>3. Descriptive Statistics for Prompts and Enhancements used Across Each Grade</td>
<td>31</td>
</tr>
<tr>
<td>4. Descriptive Statistics for Total Raw Scores and Median MAZE Scores</td>
<td>33</td>
</tr>
<tr>
<td>5. Fifth Grade Item Characteristics</td>
<td>35</td>
</tr>
<tr>
<td>6. Eighth Grade Item Characteristics</td>
<td>36</td>
</tr>
<tr>
<td>7. Biology Item Characteristics</td>
<td>37</td>
</tr>
<tr>
<td>8. Average Percentage of Disagreements per Item and Ranges of Disagreements</td>
<td>39</td>
</tr>
</tbody>
</table>
Introduction

Think Aloud / Cognitive Labs

Think aloud labs, also known as cognitive labs, require respondents to talk aloud about their cognitive processes as they are completing items or solving problems. Think aloud labs typically occur in a one-on-one environment with a researcher and a participant, and the participants are asked to think aloud, or say everything that comes to their minds as they solve problems or work through tasks (Leighton, 2017). Think aloud labs have become particularly useful to educational testing specialists for two main purposes: “to supplement empirical evidence gathered about students’ test item responses and to generate claims that educational and psychological tests measure specific constructs” (Leighton, 2017, p. 3).

Think aloud procedures help to make unobservable aspects of human cognition observable in the forms of verbal reports and behaviors as the participant interacts with the instrument. The information gathered through the think aloud procedure may be supplemented with item analysis protocols and surveys of participants. Item analysis protocols are methods used to record respondents’ thoughts directly after they complete items. Protocols may also note the item features to which respondents attend, such as visuals (e.g., charts, graphs, pictures), item passages/stimuli, item stems, answer choices, or respondents’ references to prior knowledge. This information helps to determine whether or not the test evokes the intended cognitive processes.

Peer Review Process

According to the U.S. Department of Education (2018), the peer review process “is the process through which a State documents the technical soundness of its assessment system” (p. 4). As part of the development of a technically sound state assessment system, the state must
undergo a peer review process. During this process, the state must submit evidence that demonstrates adherence to seven critical elements. The seven critical elements are as follows:“(1) Statewide System of Standards and Assessments, (2) Assessment System Operations, (3) Technical Quality – Validity, (4) Technical Quality – Other, (5) Inclusion of All Students, (6) Academic Achievement Standards and Reporting and (7) Locally Selected, Nationally Recognized High School Academic Assessments [if applicable]” (p. 29).

The think aloud lab in the current study assisted the state with the third element (Technical Quality-Validity). Validity based on cognitive processes is included within this element (Critical Element 3.2). States are advised to include particular aspects in their submission of evidence to the department. For Critical Element 3.2, states must provide evidence that their assessments evoke the intended “cognitive processes appropriate for each grade level as represented in the State’s academic content standards” (p. 49). The evidence gathered helps determine the extent assessment items align with the state academic standards, and the extent items evoke the cognitive processes intended by the test developer. Cognitive lab results indicating that “the items require complex demonstrations or applications of knowledge and skills” (p. 50) are sufficient to serve as validity evidence based on cognitive processes.

**Validity Evidence Based on Response Processes**

Validity evidence based on response processes provides information on whether or not test-takers interact with a test in the manner that was intended by the test developer (Gorin, 2006; Kettler, 2019). According to the most recent version of *Standards for Educational and Psychological Testing* (*Standards for Testing*; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education
[AERA, APA, & NCME, 2014], “validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (p. 11). Validity remains paramount for the interpretation and use of test scores. It is vital to gather appropriate evidence to support the inferences that will be drawn from test scores. One of the ways to accomplish this is through evidence based on response processes, which involves questioning a sample of the “intended test-taking population about their performance strategies or responses to particular items” (AERA, APA, & NCME, 2014, p. 15).

**Literature Review**

**Messick’s Views on Validity**

Historically, validity has been segmented into “content, criterion, and construct” validity (Messick, 1995, p. 741). Messick argued that validity should be viewed as a unified, comprehensive concept rather than as a piecemeal concept or merely as a statistic. Validity as a whole is a “social value” (p. 742), especially with regard to decision-making and forming inferences based on test performance. Construct validity is the foundation for other types of validity, and the author identified two threats to construct validity: construct underrepresentation and construct-irrelevant variance. Included in construct-irrelevant variance are “construct irrelevant difficulty” and “construct irrelevant easiness” (p. 742). Construct-irrelevant variance must be addressed and examined during “the validation process” (p. 744). In terms of educational testing, construct irrelevant difficulty involves features of the task impeding a student’s ability to demonstrate their knowledge about the target construct. To illustrate this concept, Messick gave the example of a high reading load on a test that was intended to measure math knowledge. Construct irrelevant easiness involves a tool being used to measure a certain construct that is too easy for some individuals.
Messick’s Views on Think Aloud Procedures

According to Messick (1995), one of the easiest ways to gather validity evidence may be “querying respondents about their solution processes or asking them to think aloud while responding to exercises during field trials” (p. 743). In the event that results do not work the way the researcher had hoped or intended, the process of gathering validity evidence will still yield valuable information. Although validity can be conceptualized in a general manner, there is still value in examining specific units of validity, especially considering the ways in which the scores will be used. One component of construct validity is “substantiative” (p. 745). To illustrate the substantiative aspect of construct validity, the author discussed the importance of gathering evidence that the assessment task brings the intended processes to mind. A way to accomplish gathering such evidence is through “think aloud protocols” (p. 745). Investigators are not just concerned with whether or not the task covers the domain of interest; they are also interested in whether or not the respondents engage in the intended processes.

Messick also discussed the effects associated with gathering evidence of construct validity and stated that “low scores should not occur because the measurement contains something irrelevant that interferes with the affected persons’ demonstration of competence” (p. 746). An important outcome of gathering validity evidence based on response processes is ensuring that such a situation does not occur, and that the task allows a respondent to show their knowledge of the target construct. In addition, “adverse social consequences should not be attributable to any source of test invalidity” (p. 748). Think aloud labs incorporate student voice and provide evidence that any inferences or decisions made based on these measures are cogent and will benefit students in their educational careers.
Accessibility Theory

Kettler, Braden, and Beddow (2011) discussed the interplay of test taking skills and access in educational testing. The authors defined access as “the opportunity for a student to demonstrate what she or he knows on a test” (p. 147). Construct validity refers to a test’s ability to measure the construct that it was designed to measure. According to Kettler et al., both construct relevant and construct irrelevant variance exist in testing. Construct relevant variance is related to the construct being measured, whereas construct irrelevant variance is unrelated to the target construct. The goal of a technically sound test is to reduce construct irrelevant variance in order to best capture a student’s proficiency on the target construct. The authors proposed two ways to reduce construct irrelevant variance. One way involved developing tests that do not require the use of access skills. Another way was to train students to use certain access skills, and one such skill is “test wiseness.” According to Millman, Bishop, and Ebel (1965), “test wiseness” is defined as “a subject’s capacity to use the characteristics and formats of the test and/or test taking situation to receive a high score” (p. 707). Test wiseness is not teaching to the test; it reduces barriers so that valid inferences about student performance can be drawn from the scores.

According to Kettler et al. (2011), it is important to balance teaching test-taking skills with delivering content in the classroom setting. It is beneficial for students to learn some test-taking skills; this instruction should only consume a small amount of class time and should not completely replace classroom instruction. It is important to weigh the benefits of teaching test-taking skills on a case-by-case basis according to student need. Test developers are also advised to include practice modules for computer-based tests to be delivered either during instructional time or embedded within the testing time. This will ensure that test takers are familiar with the
Universal Design

Universally designed assessments may help ensure fair and equitable access for students. According to Thompson, Johnstone, and Thurlow (2002), “universally designed assessments are not intended to eliminate individualization, but they may reduce the need for accommodations and various alternative assessments by eliminating barriers associated with the tests themselves” (p. 5). Rather than teach students ways to overcome barriers such as high reading loads, Thompson et al. proposed constructing tests carefully and including elements of Universal Design as alternative solutions.

The authors listed the elements of universally designed assessments as the following: “inclusive assessment population; precisely defined constructs; accessible, non-biased items; amenable to accommodations; simple, clear and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility” (p. 6). The second element, precisely defined constructs, reflects construct validity. According to Thompson et al., “an important function of well-designed assessments is that they measure what they actually intend to measure” (p.8). The authors argued that universally-designed assessments, by their very nature, eliminate construct-irrelevant variance.

Thompson et al. (2002) also discussed computer-based tests in relation to Universal Design. Computer-based assessments have many advantages in regard to accommodations, because they are easily programmed to incorporate text-to-speech features and on-screen reading supports, such as line highlighters. The authors emphasized the importance of familiarity with technology specific to the testing situation for students that are expected to take computer-based

computer-based test interface and can navigate it successfully, increasing the likelihood that valid inferences can be drawn from the obtained scores.
tests. Students should be given opportunities to access and become comfortable with the technology they will be expected to use prior to the testing situation.

Thompson, Johnstone, and Thurlow (2002) discussed the application of Universal Design principles to large-scale assessments as a way to optimize opportunities for all students to appropriately demonstrate their learning. Thompson et al. highlighted Universal Design in both instruction and assessment. For instructional purposes, it is insufficient to simply have lofty standards for students. Instead, it must be ensured that students can access the material during instructional delivery. Similarly, instead of trying to modify tests after they have already been constructed, the authors argued that it would be much more efficient to construct tests that are inclusive from the development phase.

Kettlerin-Geller (2008) proposed a model for examining the interaction between student characteristics and assessment tools within the Universal Design framework for students with disabilities and English learners. The author argued that principles of Universal Design should be considered at all phases of test development. The constructs of interest and the complexity of the items should not be compromised. Universal Design principles would allow students with disabilities and English learners access and the ability to demonstrate their knowledge and skills and prevent construct-irrelevant factors from interfering. The author also opined that test developers should provide the intended construct for each item so that teachers can decide which accommodations would be appropriate on an item-by-item level. This approach would allow students to demonstrate their abilities related to a given construct and overcome barriers, such as high reading loads on items testing mathematical knowledge and skills. The author also mentioned the need for flexible testing environments customized to each student’s unique needs, which could be accomplished through the use of computer adaptive tests.
**Test Development**

Gorin (2006) urged test developers to engage in a rigorous process during the design phase that is based on principles from cognitive psychology. The author argued that developing a clear definition of constructs can be achieved by using a cognitive model. In addition, Gorin encouraged test developers to use information gathered from think aloud labs and interviews with students. Taking the time in the initial stages of test development to clearly define the construct of interest, develop items that represent the domain, and test hypotheses using verbal reports will produce rich data and ensure that valid inferences may be drawn from test scores.

According to Beddow, Kurz, and Frey (2011), barriers arise during the interaction between the test taker characteristics and the test itself. The issue then becomes whether or not the test is a fair measurement of students’ abilities and knowledge. The validity of inferences and decisions based on these test scores are also questioned. The authors argued that test developers could use principles of Universal Design to ensure that a diverse range of students can participate in the same test. The authors also cautioned that it is unrealistic to expect all tests to be universally designed. Instead, under Accessibility Theory, the test should be accessible to the target population of test takers. Test developers must be familiar with the specific characteristics of the intended population. The authors also defined Ancillary Requisite Constructs (ARCs), which may interfere with the measurement of the intended construct. For example, students with visual impairments may encounter barriers when tests require them to perceive visual information in order to produce a response. The ARC in this case is the ability to process visual input, which may prohibit the students from demonstrating the knowledge and/or skills that the tests were intended to measure. Therefore, during the design of accessible tests, it is important to ensure that ARCs do not interfere with a test taker’s score and the test only measures the
intended construct. The authors also opined that test/item development and modification should be an iterative process consisting of collaboration among experts in assessment, specialists in the content areas, and staff familiar with the intended population of test takers.

The authors also illustrated test development procedures under Cognitive Load Theory (CLT). According to Beddow et al. (2011), there are three distinct categories of cognition in CLT: intrinsic (cognition necessary to complete a task); germane (cognition that helps transfer input to long term memory); and extraneous (ancillary requisite cognitive processes irrelevant to the task at hand). The authors posited that during test design under CLT, the goals are to maximize intrinsic load, reduce extraneous load, and use germane load in a purposeful manner. Test developers should ensure that these aspects are taken into account at the test and item levels.

Beddow et al. (2011) also demonstrated the use of a rubric to evaluate the accessibility of a high school science item. The authors first described an earlier version of the rubric, the Test Accessibility and Modification Inventory (TAMI; Beddow, Kettler, & Elliott, 2008). The most recent version is the TAMI Accessibility Rating Matrix (ARM; Beddow, Elliott, & Kettler, 2009), which is used to evaluate multiple choice items on the following elements: item passage/stimulus, visual, item stem, answer choices, page/item layout. Each item is rated on a four-point, Likert-type scale (Beddow, Elliott, & Kettler, 2010).

The first step consisted of the rater attempting to solve the item as a test taker and engaging in a think aloud process while solving it. The second step was to identify possible barriers in the item stimulus (e.g., the passage that precedes the item and contains related information), including high reading loads and unfamiliar or complex vocabulary. Third, the rater engaged in the same process with the item stem (e.g., the question). Fourth, the rater evaluated the visuals and whether or not they were essential to answer the question. In the event
that the visuals were necessary to answer the item, the rater also decided whether the graphs or charts were clear and easy to read. Fifth, the rater considered the answer choices. The choices should be plausible options, and some should consist of common errors. The sixth step was to evaluate the page/item layout. The raters considered working memory concerns and whether or not the layout required students to search for information. Another consideration was that an excessive amount of text at the top of a page may be daunting for struggling readers. Finally, the raters gave an overall accessibility rating, which was completed after rating each individual element. The raters also offered suggestions to improve the items.

**Information Processing Model**

According to the information processing model, the subject attends to a task, information in the short-term memory is easily accessible, and the subject is able to provide a verbal report. In the instance that a subject needed to access information stored in long-term memory, the information would be transferred to the short-term memory, and then reported aloud. The instruction to think aloud would not change the cognitive process, although it may take more time. The instruction for subjects to verbalize during task completion does not interfere with the cognitive process, nor does it require an additional cognitive load. Ericsson and Simon (1980) referred to this process as concurrent verbalization. The verbal report occurs simultaneously as the subject attends to and completes a task.

Asking subjects to explain their reasoning has an effect on the cognitive process and task completion. Similarly, changing the task does not give an accurate representation of the cognitive process required to solve a problem. For example, asking a subject to verbally instruct an experimenter regarding manipulation of a Tower of Hanoi puzzle would drastically change the demand compared to an instruction to think aloud as the subject completed the puzzle.
Subjects are generally not asked about specific processes for certain items after they have completed a large number of items. The subjects may not remember their cognitive processes, and the information they recall may be tainted by elapsed time. (Ericsson & Simon, 1980). Subjects may be asked general questions at the end of the session, such as questions regarding their thoughts about the tasks as a whole. Specific probes may limit the type of information gleaned, as they may cause the subjects to produce information that they assume the experimenter is seeking. Inconsistencies in verbal reporting may result from cues that are too general, which may lead subjects to produce information that is unrelated to the task due to unclear expectations.

Ericsson and Simon (1993) recommended that the researcher be seated behind the participant so that no social interaction would be expected. The authors also recommended that the participants be given opportunities to complete practice questions to learn the think aloud procedure. The authors recommended collecting both concurrent and retrospective data as appropriate, and they cautioned that retrospective data may be susceptible to effects of memory. The researchers also recommended that retrospective data be collected directly after the completion of an item or trial. Retrospective probes administered after the completion of multiple items over a period of time may not yield accurate responses, as participants must remember the ways they solved problems or re-solve the items to access the information.

Think Aloud Lab Theory

Historically, the field of psychology has sought to understand human processes through a variety of methods. Some methods included introspection, response times, and eye movement data. Over time, behaviorism and similar schools of thought became more widely accepted, as behavior is clearly defined and observable (Benjamin, 2007; Ericsson & Simon, 1980). Ericsson
and Simon (1980) argued that verbalizations are a type of behavior and are valuable sources of information. According to the authors, “verbal reports, elicited with care and interpreted with full understanding of the circumstances under which they were obtained, are a valuable and thoroughly reliable source of information about cognitive processes” (p. 247). The authors claimed that verbal reporting rests on a theoretical framework, similar to concepts in the physical sciences. For example, weighing an object ties into a theory about the way a pan balance works. Verbalizations give insight on the ways people produce those verbal responses. The authors proposed “a model for the verbalization processes of subjects instructed to think aloud, give retrospective reports, or to produce other kinds of verbalizations in response to experimenters’ instructions” (p. 217). The model was based on human information processing theory.

Afflerbach and Johnson (1984) suggested advantages to analyzing verbal reports. According to the authors, verbal reports possess validity based on “a different set of assumptions from those of most other methods of investigative cognitive processes” (p. 308). Verbal reports do not always align with ideas in regard to traditional hypothesis testing. This is advantageous within the context of think aloud labs, because researchers collect and analyze qualitative data based on participants’ responses. This process enables researchers to acquire valuable information in regard to test construction and the ways in which participants interact with the instrument. Other advantages to verbal reports include truthfulness that could only be ascertained from the participant directly, access to the reasoning behind their thinking to which an investigator may not be privy in traditional studies, and its availability as compared to other, more involved methods such as functional MRIs or eye tracking data. The authors also suggested that participants engage in training and practice in the verbal report process, and that researchers deliver “neutral instructions” (p. 309). The authors argued that during verbal reports, subjects are
engaged in a dual task performance: the primary, or “experimental task” and the “verbal reporting task” (p. 310). The authors also stated that answering questions allows for more “natural, frequently spaced breaks in the task” (p.310) as opposed to reading a text and being expected to verbalize while reading.

The researchers viewed concurrent and retrospective reporting as a “continuum rather than a dichotomy” (Afflerbach & Johnson, 1984, p. 311). There are strengths and weaknesses of both methods. Concurrent reporting allows one to gather information as the processing is occurring and is less likely to be tainted by time and memory. Concurrent reporting poses minimal risk to disrupt the cognitive flow. Retrospective reporting allows students to verbalize their thinking after the problem is solved or the passage is read. This method may be susceptible to time delays and may not be as accurate as live, “in the moment” data.

In regard to probing, the authors proposed that researchers should decide ahead of time which probes to provide to subjects and times at which they should be provided. Probes such as “keep talking” or “keep thinking aloud” fit into Afflerbach and Johnston’s (1984) suggestions since they avoid inadvertently prompting the student to think a particular way, which may interfere with the cognitive process. The selection of subjects is dependent on the purpose of the study and the research questions.

It is also important to find the questions and activities at the appropriate difficulty level. Items should not be too difficult, as this would lead students to struggle or cease verbalization. Items should not be too easy, as this would display automaticity (Afflerbach & Johnson, 1984; Ericsson and Simon 1980). Ideally, the items should be at an intermediate level in order to access problem solving strategies and provide meaningful information. During analysis of the response data, Afflerbach and Johnson (1984) suggested that one should consider the features of spoken
language, including “intonation, inflection, pauses, and variation in rate of speech” (p. 315). In addition, one should not be overly rigid or have a specific number of utterances in mind from the outset. Instead, the researcher should allow space for individual differences and novel strategies and responses. The authors also suggest mapping “solution paths” (p. 317) to show the strategies a student used.

The authors mention the importance of training raters to increase interrater reliability. In regard to research design, Afflerbach and Johnson (1984) agree that these types of studies should involve a small number of subjects due to the labor and time intensive nature of the analyses. In addition, the tasks should be untimed. The authors suggest the use of “multiple indications” (p. 319) for the analysis of responses. The authors stated that in the event alternative methods are unavailable (e.g., eye movement data, fMRI), concurrent and retrospective accounts are sufficient to be considered multiple data points. According to the authors, “used appropriately, verbal reports offer a unique, if sometimes less than transparent, window for viewing cognitive processes” (p. 320). Although psychometric data exists that attempts to quantify invisible constructs, the authors support the notion that verbal reports are as valid as any other measure. The responses come directly from the source in which the cognitive process occurs.

In a think aloud situation, students’ perceptions that raters are “experts” in the content area could affect students’ response accuracy (Leighton, 2013). Similarly, Ericsson and Simon (1998) posited that “the first step toward studying covert thinking requires that we find a nonreactive setting to reproduce this type of thinking under controlled conditions” (p. 179). Ericsson and Simon argued that in order to evoke an accurate think aloud, the setting must be non-evaluative and non-judgmental. The authors proposed a model that illustrated the way one’s “silent thinking” is identical in process to “thinking aloud” (p. 180). Differences arise in the
event one is asked to describe or explain their thinking or problem-solving process. According to the authors, it is crucial for the participant to maintain focus on the primary task (e.g., answering a science question) rather than be asked to explain their thinking. The authors also stated that the thoughts verbalized during the think aloud process are not likely to be linear or cohesive. During analysis of think aloud protocols, one is likely to encounter incomplete phrases and other indicators of internal thought. The authors also stated that intercoder reliability is an utmost priority during think aloud protocol analysis.

According to Ericsson and Simon (1993), “concurrent verbalization” (p. xiii) can be attained by instructing a participant to “think aloud” or “keep talking.” The authors differentiated between three types, or levels, of verbalizations. The first level, Type 1 verbalizations, occur while a participant is completing a task. According to the authors, concurrent verbalizations while solving a problem are the closest to inner thoughts and dialogue. Verbalizations produced at this level may be incoherent and disjointed; asking a participant to verbalize at this level does not interfere with their cognitive process. The authors stated that there may be slight differences between the verbalizations that participants make to themselves, versus the utterances produced for others to hear.

The second level, Type 2 verbalizations, are similar to Level 1 in that they do not require the participant to retrieve new information. Rather, Level 2 verbalizations involve a participant describing their thinking on a particular item. Retrospective verbal reports obtained after the completion of items are considered Type 2 verbalizations.

The third level, Type 3 verbalizations, involve a participant explaining their thinking or solutions. The demands to achieve Type 1 and Type 2 verbalizations are different than the instructions given to achieve Type 3 verbalizations. Type 1 verbalizations can be attained by
instructing participants to “think aloud,” and Type 2 verbalizations can be ascertained by asking an open-ended question about the participant’s thought process directly after the completion of an item. Type 3 verbalizations usually follow a prompt for the participants to “explain” their thinking. This instruction would require one to retrieve information from memory, encode a response, and offer a higher level of evaluation.

**Think Aloud Studies**

There are several think aloud studies in the literature that were used to develop the rationales and procedures for the current study. Table 1 summarizes the specifications of each think aloud study.

**Table 1**

*Specifications of Think Aloud Studies of Achievement Tests*

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Content Area</th>
<th>Forms</th>
<th>Items per Form</th>
<th>Groups</th>
<th>Students</th>
<th>Time</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dickenson et al.</td>
<td>2011</td>
<td>Science</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>18</td>
<td>NR</td>
<td>Yes</td>
</tr>
<tr>
<td>Erickan et al.</td>
<td>2010</td>
<td>Science</td>
<td>2</td>
<td>20</td>
<td>2</td>
<td>48</td>
<td>45 min</td>
<td>No</td>
</tr>
<tr>
<td>Johnstone et al.</td>
<td>2006</td>
<td>Science</td>
<td>1</td>
<td>12</td>
<td>5</td>
<td>85</td>
<td>NR</td>
<td>Yes</td>
</tr>
<tr>
<td>Lau, K.</td>
<td>2006</td>
<td>Reading</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>8</td>
<td>30-50 min</td>
<td>No</td>
</tr>
<tr>
<td>Leighton, J.P.</td>
<td>2013</td>
<td>Math</td>
<td>1</td>
<td>15</td>
<td>3</td>
<td>71</td>
<td>45-60 min</td>
<td>No</td>
</tr>
<tr>
<td>Roach et al.</td>
<td>2010</td>
<td>Reading and Math</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>9</td>
<td>20-30 min</td>
<td>Yes</td>
</tr>
<tr>
<td>Winter et al.</td>
<td>2006</td>
<td>Math</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>156</td>
<td>45 min</td>
<td>No</td>
</tr>
</tbody>
</table>

*Note.* NR = Not recorded; Acc = Accommodations.
Erickan et al. (2010)

Erickan, Arim, Law, Domene, Gagnon, and Lacroix (2010) used think aloud protocols to investigate Differential Item Functioning (DIF) between two groups of students (English-speaking and French-speaking). Experts analyzed two forms of an assessment, an English version and a French version, to determine which aspects about certain items would either be helpful or confusing to students. The authors argued that relying on expert analysis is not sufficient for determining which aspects of an item are either helpful or confusing for students while they are solving problems on assessments. Research to gather data related to the students’ experiences is necessary to understand their experiences of working through the problems.

The researchers recruited two groups of students. One group consisted of 36, English-speaking seventh grade students from two schools in Vancouver. The second group consisted of 12 French-speaking seventh and eighth grade students from five schools in a French-speaking school district in Vancouver. The 20 test items included 12 math problems and eight science questions. Nine questions were presented in a multiple-choice format, and 11 questions were presented in an open-ended format and required a short, written response. The test administrators were trained to ensure standardized proctoring practices. Test administrators were asked to record the times the students spent on each item, as well as any observations, and to administer planned prompts throughout the testing sessions. The sessions were conducted in a one-on-one scenario and lasted approximately 45 minutes. Student responses were recorded, transcribed, coded, and analyzed for common categories of responses. The test administrators were unaware of the differences between the tests, to avoid possible bias.

Four categories emerged from the students’ responses: understanding the question; perceived difficulty of the question; features that helped the students solve the problems; and features that hindered the students from solving the problems. Results revealed that English-
speaking students performed better than French-speaking students on their respective assessments. For ten of the items, the think aloud responses corroborated “linguistic differences as sources of DIF” (Erickan et al., 2010, p. 28). The researchers found evidence of “linguistic differences not identified by expert reviewers” for one of the items (p. 28), and for the remaining nine items, the think aloud responses did not support the aspects that the expert reviewers had identified as potential “sources of DIF” (p. 28). In some instances, students’ verbalizations brought issues to light that experts had not identified in their analyses. In other cases, the students’ verbalizations either corroborated or refuted the findings highlighted in the experts’ analyses.

Erickan et al. (2010) recommended that during a think aloud, multiple sources of evidence should be combined, including the number of items that students answered correctly and incorrectly, the categories of responses that emerged, time it took for the students to complete each item, and any observations that the test administrators noted during the process. The three categories that the researchers found to be the most informative were students’ understanding of the questions, features that helped the students solve the problems, and features that prevented them from solving the problems. According to the authors, information on whether or not an item is difficult is not helpful, since students can misinterpret difficult aspects of an item. Many students found certain items hard and still answered them correctly. The authors also support the use of both concurrent and retrospective reporting. According to the researchers, retrospective verbal data should only consist of students’ general reactions to the problems.
Dickenson et al. (2011)

Dickenson, Gilmore, Price, and Bennett (2011) used the think aloud method to evaluate state-wide, high school science test items that were to be used on alternate assessments in South Carolina. Eighteen students participated in total, representing three groups: students without disabilities (SWOD); students with disabilities that were eligible for alternate assessment (SWD-E); and students with disabilities that were not eligible for alternate assessment (SWD-NE). Dickenson et al. (2011) selected six items, three of which were stand-alone items and three of which were based on a passage. The researchers administered two different versions of the tests, which meant that one group completed stand-alone items in either the enhanced or original version, and another group completed passage-based items, either enhanced or original. Using theories of Universal Design for Learning and Cognitive Load, the authors chose to offer the following item enhancements to certain questions: an eliminated answer choice; an added graphic; simplified vocabulary; a shortened stem; added white space; and having the item read aloud. The students were taught to think aloud and practiced the procedure using lower grade-level items. Students were videotaped and asked to think aloud as they answered each question. The researchers prompted the students to “keep thinking aloud” or “keep talking” in the event they stopped verbalizing for three consecutive seconds. After the think aloud lab portion, the researchers administered follow-up questionnaires and oral reading fluency measures.

Results indicated that oral reading fluency rate scores were highest for SWOD and lowest for SWD-E. There was a statistically significant positive correlation between oral reading fluency rate scores and raw scores on the original items. The correlation between oral reading fluency and raw scores for the enhanced items was not statistically significant. For SWOD, the raw scores on the enhanced items were slightly higher, and for both SWD-E and SWD-NE, there
were similar total raw scores for original and enhanced items. All students reported lower perceived cognitive effort on enhanced items as compared to original items. There was no difference in time spent solving items for SWOD; both SWD-E and SWD-NE spent more time on the original items. SWD-E performed better on stand-alone items than passage-based items. This group of students also had the largest mean difference in time spent solving passage-based versus stand-alone items.

Students reported that the enhancements helped them to understand the items, including in particular the use of visuals, removal of a distractor, and the read-aloud feature. Students reported that the visuals were helpful on stand-alone items more frequently than on passage-based items. Dickenson et al. (2011) concluded that passage-based items may be more difficult for SWD-E, and this group may require additional enhancements. Although the SWD-E performances were not improved with the use of enhancements, the students perceived these features as helpful, and they reported decreased levels of cognitive effort and spent less time solving the items.

**Roach et al. (2010)**

Roach, Beddow, Kurz, Kettler, and Elliott (2010) used the think aloud method to evaluate items on two versions of eighth grade reading and math tests. Both versions contained eight items total: four modified and four unmodified items. Nine eighth-grade students represented three groups of interest: students with disabilities eligible for alternate assessment (SWD-E), students with disabilities not eligible for alternate assessment (SWD-NE), and students without disabilities (SWOD). Each student completed both a math test and a reading test. As outlined by Johnstone et al. (2006), students were introduced to the think aloud procedure, the researchers modeled the task, and the students completed practice items. Similar to Johnstone et al. (2006),
the researchers prompted the students after 10 consecutive seconds of silence by stating, “Keep thinking aloud” or “Keep talking.” Roach et al. (2010) asked the students follow-up questions about their responses and/or about the items. The researchers reported the following data after the students completed the tests: percentage of items correct, average amount of time spent on each item, average number of miscues on reading passages, fluency rate on passages (words correct per minute) and the number or times the researcher prompted the student per item (Roach et al., 2010, p. 69). The researchers asked students follow-up questions regarding the following modifications: visuals and graphics, bold font with key words or vocabulary, reduced number of answer choices, and alternative analogy formats (pp. 69-70). The researchers administered post-test questionnaires in a separate study, which involved a large number of students from several states.

Overall, the students reported that the modifications were helpful, including in particular the reading support feature, pictures and visuals, and decreased number of answer choices. The modified version of the test seemed easier to the students, and the modified graph was easier for students to understand. Roach et al. (2010) reported that reading fluency was a barrier; for some SWD, it took almost twice as long to complete the test as compared to SWOD. The researchers found that for all students, the modified version of the test took less time to complete and required fewer prompts during the think aloud procedure.

**Johnstone et al. (2006)**

Johnstone, Bottsford-Miller, and Thompson (2006) used the think aloud method to assess the design of state-wide science assessments in Minnesota for elements of Universal Design for Learning (UDL), as outlined in Thompson, Johnstone, and Thurlow (2002). The authors attempted to obtain five to 10 students from each of the following five categories in both fourth
and eighth grades: students with learning disabilities, English language learners, students that were deaf or hard of hearing, students with cognitive disabilities, and students without disabilities that were English language proficient. The students completed 12 test items: six from the fourth-grade test and six from the eighth-grade test. The questions were chosen based on past data, which revealed that the majority of students with learning disabilities and English language learners answered them incorrectly the previous year.

Johnstone, et al. (2006) used a two-part process by which verbal data was recorded as the students solved problems, and the researchers asked follow-up questions following the completion of the items. The researchers developed a script in which they explained the process to the students, gave the students opportunities to ask questions, modeled the think aloud process, and allowed the students to practice thinking aloud using a hidden picture task. Students were informed that they would be videotaped. In the event the students fell silent for 10 consecutive seconds during the think aloud process, the researchers prompted the students by saying, “Keep talking” or “Keep thinking aloud.” The researchers did not have a definitive set of questions for the follow-up session; questions were related to occurrences that arose as the students answered the items. Some questions were process-based, and others were related to the test design. The researchers created a coding sheet which had areas to record information regarding students’ test taking behaviors. The students’ written responses (e.g., whether or not the student answered correctly) and verbal behaviors were coded from the videotapes. Finally, using the coded results from the students’ verbal responses, the researchers evaluated the test items for Universal Design elements (Thompson, et al., 2002).

Johnstone, et al. (2006) reported their results in terms of UDL elements. The think aloud process revealed problems with Element 2, Precisely Defined Constructs, for three items. One
item required a two-part solution. Students did not attend to that part of the question and were unable to correctly apply their mathematics skills. Another item required students to use a map, and the directions were unclear to students. Students expressed confusion about which calculations to perform and the utility of the map. Students also expressed confusion in regard to the demand of a third item, and they reported being distracted by visuals. For Element 3, Accessible, Non-biased items, issues were detected due to students’ inexperience with a certain type of question. Think aloud procedures indicated that for Element 5, Simple, Clear, and Intuitive Instructions, some students answered certain items incorrectly due to unclear or confusing instructions. For Element 6, Maximum Readability and Comprehensibility, students reported difficulty with vocabulary, and a student with a hearing impairment was confused by a word with a double meaning. Finally, for Element 7, Maximum Legibility, issues were detected because students read two numbers incorrectly that looked very similar. In addition, the computer misaligned a fraction, which confused a student.

**Winter et al. (2006)**

Winter, Kopriva, Chen, and Emick (2006) used cognitive lab procedures to investigate the interaction between individual factors and item factors, and particularly the ways in which the interaction affected access. In a sample of 156 students, the researchers gathered information about each student, including cultural background, level of English proficiency, achievement levels, and cognitive ability. The authors matched the students with math items that were designed to reduce barriers to access according to their individual needs. The students were also administered typical, grade-level items. Winter et al. used a problem-solving model that involved assessing whether or not the students understood the demand of the question, chose an appropriate strategy, applied the strategy to solve the problem, and articulated their responses.
During the cognitive lab, the researchers recorded their observations according to standardized protocols. Following the completion of the items, students were administered retrospective interviews, the students’ responses were coded and their answers to the test items were scored.

The results indicated that individual factors interacted with construct-irrelevant factors. This outcome supports the authors’ hypothesis that the interaction affects access. Matching items with individual students’ needs reduced the barriers, and the students were able to understand the task demands. The students that understood the task were more likely to choose the correct solution strategy and solve the problem correctly. Winter et al. (2006) argued that their findings have implications for assessment validity, in particular: “To the extent that students face difficulty with the construct-irrelevant requirements of an item, the item becomes less able to provide accurate information about their achievement in the area of the targeted outcome” (p. 267). Students are not able to “show what they know” if their access is impeded with barriers.

**Current Study**

The current study helped to evaluate whether the testing items evoked the intended cognitive and problem-solving processes. In addition, this study examined whether or not the items were valid ways for students to demonstrate their knowledge in science, barriers that may have impeded their access (e.g., high reading loads), and whether or not valid inferences could be drawn from the results in regard to students’ knowledge of science content.

The current study sought to gather evidence based on response processes during a think aloud lab which followed a protocol that Johnstone, Bottsford-Miller, and Thompson (2006) developed.

**Research Questions and Predictions**

The think aloud lab will answer three primary research questions:
1.) What does the validity evidence based on cognitive processes indicate with regard to the new Tennessee state science assessments?

2.) How do the questions on state science assessments evoke intended cognitive processes?

3.) What is the relationship between Maze scores and number of correct responses to science items?

It was hypothesized that the validity evidence based on cognitive processes would indicate that the new Tennessee state science assessments are valid measures of student knowledge of science according to the state standards for each grade level. It was also hypothesized that the questions would evoke intended cognitive processes through reading passages and referring to visuals (e.g., charts, graphs). Finally, it was hypothesized there would be no significant relationship between number of items answered correctly and scores on the Maze reading probes.

Method

Participants

Participants were recruited by the Tennessee State Department of Education. The sample was one of convenience based on proximity to the city and willingness to participate. According to Nielsen (1994), five subjects is sufficient to provide enough information for using the think aloud method to detect design or usability issues.

The students represented three grades (fifth, eighth, and high school) and the students that were recruited were relatively high performing. Higher performing students were necessary because thinking aloud creates an additional cognitive load, making the task challenging for lower performing students. Higher performing students would be able to complete the think aloud process and provide meaningful data.
Students from two school districts participated in think aloud lab sessions. According to the National Center for Educational Statistics (2019), both White House Heritage schools in Robertson County and Smyrna School District in Rutherford County represent large, suburban locales. The most common ethnicity for both districts is White, non-Hispanic. Both districts also have populations of Black, Hispanic or Latino, and Asian students. Approximately one in five families qualifies for food stamps or supplemental nutrition assistance program (SNAP) benefits in both districts.

Participants included students \( n = 33 \) from three grade bands (13 fifth grade students, 12 eighth grade students, and eight high school students). Table 2 summarizes demographic information of the sample by grade.

**Table 2**

Demographics of the Sample by Grade Band

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>Female</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>Other</th>
<th>FRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fifth</td>
<td>13</td>
<td>5 (38%)</td>
<td>9 (69%)</td>
<td>3 (23%)</td>
<td>1 (8%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>2 (15%)</td>
</tr>
<tr>
<td>Eighth</td>
<td>12</td>
<td>7 (58%)</td>
<td>9 (75%)</td>
<td>1 (8%)</td>
<td>1 (8%)</td>
<td>1 (8%)</td>
<td>0 (0%)</td>
<td>4 (33%)</td>
</tr>
<tr>
<td>H.S.</td>
<td>8</td>
<td>7 (88%)</td>
<td>6 (75%)</td>
<td>0 (0%)</td>
<td>1 (13%)</td>
<td>1 (23%)</td>
<td>0 (0%)</td>
<td>2 (25%)</td>
</tr>
<tr>
<td>All</td>
<td>33</td>
<td>19 (58%)</td>
<td>24 (73%)</td>
<td>4 (12%)</td>
<td>3 (9%)</td>
<td>2 (6%)</td>
<td>0 (0%)</td>
<td>8 (24%)</td>
</tr>
</tbody>
</table>

*Note. FRL = Free or Reduced Lunch; H.S. = High School.*

**Measures**

**Maze Reading Probes**

Students completed three grade-level Maze reading assessments following the think aloud sessions (Florida Center for Reading Research, 2009; Fuchs, 2017). These assessments were used to measure reading fluency and comprehension. Every seventh word from the passage was missing, and the students were instructed to circle one word that best fit the sentence out of three options. The students were asked to complete as many blanks as possible. Maze
assessments have been found to have acceptable Cronbach’s alphas for both SWD (alternate form = .77, odd/even = .92) and SWOD (alternate form = .89, odd/even = .95).

**TNReady Science**

TNReady Science tests assess the Tennessee Academic Standards, which require students to demonstrate deep conceptual understanding across three dimensions: disciplinary core ideas, science and engineering practices, and crosscutting concepts. The fifth, eighth, and high school tests consist of 45 items to be completed in 75 minutes via an online platform. The online testing program featured the following standard accessibility tools: bookmark, note, highlighter, answer eliminator, and line reader. Participants that took part in the think aloud labs did not have access to any other accommodations beyond the standard accessibility tools. Each of the tests consisted of both cluster set items and discrete items.

**Cluster Set Items**

Each cluster set item is designed around a common phenomenon and is meant to reflect the complexity of the Tennessee Academic Standards for Science. The items in the set share common stimulus material (e.g., passages, charts, tables, or images), and the items are scored separately. Both multiple choice and multiple select questions were part of cluster set items. Multiple choice items require the students to choose one response, while multiple select items required each student to choose more than one response.

**Discrete Items**

Discrete items are either multiple choice or multiple select items that are not connected to any other items on the assessment. Discrete items are aligned with the disciplinary core ideas and may or may not also measure science and engineering practices or crosscutting concepts.
The current study used samples of the aforementioned items on forms of the TNReady tests. The fifth grade form consisted of 12 items (eight cluster, two multiple choice, two multiple select), the eighth grade form consisted of 11 items (seven cluster, two multiple choice, two multiple select), and the high school form consisted of 12 items (eight cluster, two multiple choice, two multiple select). The test vendor (ETS; www.ETS.org) provided rationales for the correct answer choices and the incorrect answer choices (i.e., distractors) for each item. Rationales for the correct answers could be divided into two categories: use of prior knowledge or use of the visual (e.g., table, chart, figure).

Survey

Following each test, proctors orally administered a five-item survey. The purpose of the survey was to gather student impressions of the tests and their preferences for either computer-based or paper-pencil tests.

Procedures

Following a protocol that Johnstone, Bottsford-Miller, and Thompson (2006) developed, students were introduced to the think aloud procedure with a script, were provided with models of persons thinking aloud during testing and were administered practice items to gain familiarity and comfort with the think aloud process. Both the practice items and testing items were administered to students by proctors in one-on-one environments. The proctors were four persons with advanced education in assessment trained in school psychology or employed by the Tennessee Department of Education. Following the guidance of Leighton (2013), it was made clear to the students the researchers were not experts in science and simply interested in student strategies for solving problems. Students’ perceptions that proctors were “experts” in the content area could have negatively affected response accuracy. Following the introduction, modeling
and practice items, students were asked whether they had any questions about the process. The students were asked permission to audio record the session.

The students were given unlimited time to complete the tests. Following the recommendations from Johnstone, Bottsford-Miller, and Thompson (2006), and Roach, Beddow, Kurz, Kettler, and Elliott (2010), students were prompted to “Keep talking” or “Keep thinking aloud” during periods of silence lasting about 10 consecutive seconds. The think aloud process allows collection of concurrent verbal responses as the students solve problems (Ericsson & Simon, 1980). After each think aloud procedure was completed, interview surveys were administered individually to each student to gather additional data about their response processes (Ericsson & Simon, 1980; Roach, Beddow, Kurz, Kettler, & Elliott, 2010).

Following the procedures of Roach et al. (2010), students’ verbal and non-verbal behaviors were recorded during the think aloud sessions on a coding sheet, and were also audio-recorded, transcribed, and coded. Proctors were trained to use the coding sheets consistently through multiple practice opportunities. Coding sheets were unique to each item and included the following categories: prompts used; tools used; reading and re-reading of directions, passages, and questions; and commenting on visuals. The audio recordings were transcribed by the Tennessee Department of Education using Otter online software (www.otter.ai). The transcriptions were coded and analyzed to determine which strategies were used to complete each item. The proctors converged on nine strategies used in varying frequencies across individuals and items. Following this coding, the item specifications were reviewed and it was determined that each item was designed to be solved using either prior knowledge or examination of the item visual.
Data Analysis

Data analysis was conducted using SPSS Statistics version 26.0 and Microsoft Excel. Basic descriptive statistics, such as means and standard deviations of prompts and enhancements used across each grade, were used to analyze the data. Percentages of students that engaged in each strategy were calculated from cover sheets and coded transcripts of each item. One-tailed Pearson correlations were conducted to examine the relationships between Maze scores and total raw scores on TNReady items for each grade. Two-tailed, independent sample T-tests were conducted to compare mean word counts and Flesch-Kincaid grade reading levels of testing items.

Average percentages of disagreements in transcript coding for each item were calculated in Excel. Data from the transcripts were coded by two of the proctors. The coding categories were collapsed to indicate for each individual completing each item whether prior knowledge was invoked, whether the visual was examined, or whether both strategies were used. Data sheets from the two proctors were combined in Excel and items that did not match were considered a disagreement. The total number of disagreements per item were divided by the number of students in each grade. Student responses to the post-test survey were analyzed qualitatively and grouped thematically.

Results

Prompts and Enhancements

The means, standard deviations, and minimum and maximum values of prompts used with and without cover sheets, as well as enhancements used with and without cover sheets across each grade are presented in Table 3. Cluster set items and certain discrete items were accompanied by cover sheets, which consisted of passages and/or visuals related to the
questions. As indicated in Table 3, more prompts were utilized with fifth grade students on items that included cover sheets ($M = 4.38$, $SD = 4.17$) than with eighth grade students ($M = 1.43$, $SD = 1.30$) and high school students ($M = 0.88$, $SD = 1.00$) on similar items.

High school students were more likely to utilize enhancements both on items with cover sheets ($M = 2.40$, $SD = 0.92$) and without cover sheets ($M = 2.00$, $SD = 0.60$). Eighth grade students utilized some enhancements on items with cover sheets ($M = 0.60$, $SD = 1.00$) and without cover sheets ($M = 0.20$, $SD = 0.40$), and less often than high school students. Fifth grade students seldom utilized enhancements on both items with cover sheets ($M = 0.13$, $SD = 0.40$) and without cover sheets ($M = 0.10$, $SD = 0.30$).

Table 3

*Descriptive Statistics for Prompts and Enhancements Used Across Each Grade*

<table>
<thead>
<tr>
<th>Variable</th>
<th>$M$</th>
<th>$SD$</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fifth Grade</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prompts Used with Cover Sheet</td>
<td>4.38</td>
<td>4.17</td>
<td>1.00</td>
<td>12.00</td>
</tr>
<tr>
<td>Prompts Used without Cover Sheet</td>
<td>2.42</td>
<td>2.71</td>
<td>0.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Enhancements Used with Cover Sheet</td>
<td>0.13</td>
<td>0.40</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Enhancements Used without Cover Sheet</td>
<td>0.10</td>
<td>0.30</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Eighth Grade</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prompts Used with Cover Sheet</td>
<td>1.43</td>
<td>1.30</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Prompts Used without Cover Sheet</td>
<td>0.82</td>
<td>1.10</td>
<td>0.00</td>
<td>3.00</td>
</tr>
<tr>
<td>Enhancements Used with Cover Sheet</td>
<td>0.60</td>
<td>1.00</td>
<td>0.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Enhancements Used without Cover Sheet</td>
<td>0.20</td>
<td>0.40</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>High School</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prompts Used with Cover Sheet</td>
<td>0.88</td>
<td>1.00</td>
<td>0.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Prompts Used without Cover Sheet</td>
<td>0.60</td>
<td>0.90</td>
<td>0.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Enhancements Used with Cover Sheet</td>
<td>2.40</td>
<td>0.92</td>
<td>1.00</td>
<td>4.00</td>
</tr>
<tr>
<td>Enhancements Used without Cover Sheet</td>
<td>2.00</td>
<td>0.60</td>
<td>1.00</td>
<td>3.00</td>
</tr>
</tbody>
</table>
Raw Scores and Maze Scores

The means, standard deviations, minimum and maximum values of total raw scores on TNReady science items and median Maze scores across each grade are presented in Table 4. As indicated in Table 4, mean raw scores on science items were similar across grades. The mean raw scores were slightly higher among fifth grade students ($M = 7.15, SD = 1.21$) than among eighth grade students ($M = 6.75, SD = 1.91$), and were slightly higher among eighth grade students than among high school students ($M = 6.25, SD = 3.15$).

The Maze assessment was administered in late April. Median Maze scores tended to be higher in higher grades as compared to lower grades. The highest median Maze scores were among high school students ($M = 39.75, SD = 11.41$). This average is approximately $3/4$ standard deviation above the average (Florida Center for Reading Research, 2006) for high school students in the winter ($31+$) and slightly higher than $1/3$ standard deviation above the average in the spring ($34+$). Eighth grade students restored an average of $28.25 (SD = 8.45)$ words correctly on the Maze. This average is about one standard deviation above the average for eighth grade students in the winter ($20$) and about $1/3$ standard deviation above the average in the spring ($25$). Fifth grade students restored an average of $20.35 (SD = 6.09)$ words correctly on the Maze. This average is between averages for fifth grade students (Hosp, Hosp, & Howell, 2007) in the winter ($20$) and spring ($24$).
Table 4

*Descriptive Statistics for Total Raw Scores and Median Maze Scores*

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fifth Grade</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Raw Scores</td>
<td>7.15</td>
<td>1.21</td>
<td>5.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Median Maze Scores</td>
<td>20.35</td>
<td>6.09</td>
<td>11.00</td>
<td>34.00</td>
</tr>
<tr>
<td><strong>Eighth Grade</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Raw Scores</td>
<td>6.75</td>
<td>1.91</td>
<td>3.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Median Maze Scores</td>
<td>28.25</td>
<td>8.45</td>
<td>15.00</td>
<td>42.00</td>
</tr>
<tr>
<td><strong>High School</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Raw Scores</td>
<td>6.25</td>
<td>3.15</td>
<td>1.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Median Maze Scores</td>
<td>39.75</td>
<td>11.41</td>
<td>22.00</td>
<td>55.00</td>
</tr>
</tbody>
</table>

**Student Strategy Use**

The majority of students across all three grades read the passages, questions, and directions. Students in fifth grade were less likely to access prior knowledge when the test vendor named this as a critical strategy (59%), than eighth grade students (78%) and high school students (75%). Fifth grade students were more likely to refer to a visual (93%) than eighth grade students (89%), and eighth grade students were slightly more likely than high school students to refer to a visual (83%) whenever the test vendor named this as a critical strategy. On average, fifth grade students spent more time completing items (about two to four minutes) than eighth grade students (about one to two and a half minutes) and high school students (about one and a half to three minutes).
**Fifth Grade**

Fifth grade students spent averages of about two minutes to about four minutes completing each item. All students read item passages, questions, and answers on all items. Prior knowledge was explicitly accessed by an average of 50% of students across items. Returning to a visual (figure or table) was explicitly referenced by an average of 74% of students across items. For three of the 12 items, accessing the visual was identified as a critical strategy by the test vendor. Returning to the visual was more common on average for these three items (93%) than it was for the other nine items (65%). Table 5 depicts characteristics of fifth grade items.

**Eighth Grade**

Eighth grade students spent averages of about one minute to about two and half minutes completing each item. Almost all students read item passages, questions, and answers on all items. One passage was skipped by one student and another was skipped by two students, possibly employing a test-taking strategy of going to the questions first. This makes sense under some timed test conditions; time limits were not used in the current study. Prior knowledge was explicitly accessed by an average of 65% of students across items. Returning to a visual was explicitly referenced by an average of 59% of students across items that featured a visual. For six of the 11 items, accessing the visual was identified as a critical strategy by the test vendor. Returning to the visual was more common for these six items (89%) than it was for the other five items (7%). Table 6 depicts characteristics of eighth grade items.

**High School Students**

High school students spent averages of about one and a half minutes to about three minutes completing each item. Almost all students read item passages, questions, and answers on all items. On one item, one student skipped the passage. Prior knowledge was explicitly accessed
by an average of 75% of students across items. Returning to a visual was explicitly referenced by
an average of 83% of students across items that involved at least one visual. Table 7 depicts
characteristics of biology items.

**Table 5**

*Fifth Grade Item Characteristics*

<table>
<thead>
<tr>
<th>Item</th>
<th>Standard Code</th>
<th>Rationale</th>
<th>Mean Time</th>
<th>Pass Read</th>
<th>Ques Read</th>
<th>Ans Read</th>
<th>PK</th>
<th>RTV</th>
</tr>
</thead>
<tbody>
<tr>
<td>#9</td>
<td>5.PS1.1</td>
<td>Visual</td>
<td>171s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>60%</td>
<td>80%</td>
</tr>
<tr>
<td>#13</td>
<td>5.ESS1.6</td>
<td>Knowledge</td>
<td>110s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>50%</td>
<td>33%</td>
</tr>
<tr>
<td>#3</td>
<td>5.LS3.1</td>
<td>Knowledge</td>
<td>137s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>78%</td>
<td>67%</td>
</tr>
<tr>
<td>#25</td>
<td>5.PS2.1</td>
<td>Knowledge</td>
<td>143s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>73%</td>
<td>20%</td>
</tr>
<tr>
<td>#28</td>
<td>5.PS1.2</td>
<td>Knowledge</td>
<td>152s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>25%</td>
<td>100%</td>
</tr>
<tr>
<td>#29</td>
<td>5.PS1.2</td>
<td>Knowledge</td>
<td>147s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>20%</td>
</tr>
<tr>
<td>#30</td>
<td>5.ESS1.1</td>
<td>Knowledge</td>
<td>131s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>55%</td>
<td>100%</td>
</tr>
<tr>
<td>#31</td>
<td>5.ESS1.1</td>
<td>Knowledge</td>
<td>128s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>67%</td>
<td>83%</td>
</tr>
<tr>
<td>#37</td>
<td>5.LS4.1</td>
<td>Knowledge</td>
<td>139s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>75%</td>
<td>64%</td>
</tr>
<tr>
<td>#38</td>
<td>5.LS4.1</td>
<td>Knowledge</td>
<td>109s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>9%</td>
<td>100%</td>
</tr>
<tr>
<td>#39</td>
<td>5.LS4.1</td>
<td>Visual</td>
<td>237s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>9%</td>
<td>100%</td>
</tr>
<tr>
<td>#40</td>
<td>5.LS1.1</td>
<td>Visual</td>
<td>118s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>17%</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>144s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>50%</td>
<td>74%</td>
</tr>
</tbody>
</table>

*Note.* Pass = Passage; Ques = Question; Ans = Answer; PK = Prior Knowledge; RTV = Return
to Visual.
### Table 6

**Eighth Grade Item Characteristics**

<table>
<thead>
<tr>
<th>Item</th>
<th>Vendor</th>
<th>Cover Sheet</th>
<th>Transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Item Standard Code</strong></td>
<td><strong>Rationale</strong></td>
<td><strong>Mean Time</strong></td>
<td><strong>Pass</strong></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#13</td>
<td>8.ESS3.1</td>
<td>Knowledge</td>
<td>93s</td>
</tr>
<tr>
<td>#23</td>
<td>8.PS2.4</td>
<td>Visual</td>
<td>140s</td>
</tr>
<tr>
<td>Multiple Select</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#15</td>
<td>8.PS2.3</td>
<td>Knowledge</td>
<td>98s</td>
</tr>
<tr>
<td>#24</td>
<td>8.PS2.2</td>
<td>Knowledge</td>
<td>104s</td>
</tr>
<tr>
<td>Cluster #1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#39</td>
<td>8.PS4.1</td>
<td>Visual</td>
<td>108s</td>
</tr>
<tr>
<td>#40</td>
<td>8.ESS1.1</td>
<td>Visual</td>
<td>97s</td>
</tr>
<tr>
<td>#41</td>
<td>8.ESS1.1</td>
<td>Visual</td>
<td>105s</td>
</tr>
<tr>
<td>Cluster #2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#42</td>
<td>8.PS2.2</td>
<td>Knowledge</td>
<td>66s</td>
</tr>
<tr>
<td>#43</td>
<td>8.ESS1.2</td>
<td>Knowledge</td>
<td>64s</td>
</tr>
<tr>
<td>#44</td>
<td>8.ESS1.2</td>
<td>Visual</td>
<td>81s</td>
</tr>
<tr>
<td>#45</td>
<td>8.ESS1.2</td>
<td>Visual</td>
<td>79s</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>94s</td>
</tr>
</tbody>
</table>

*Note.* Pass = Passage; Ques = Question; Ans = Answer; PK = Prior Knowledge; RTV = Return to Visual.
Table 7

*Biology Item Characteristics*

<table>
<thead>
<tr>
<th>Item</th>
<th>Standard Code</th>
<th>Rationale</th>
<th>Mean Time</th>
<th>Pass</th>
<th>Ques</th>
<th>Ans</th>
<th>PK</th>
<th>RTV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Multiple</td>
<td>Choice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#18</td>
<td>BIO1.LS1.5</td>
<td>Knowledge</td>
<td>145s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>75%</td>
<td>88%</td>
</tr>
<tr>
<td>#21</td>
<td>BIO1.LS1.9</td>
<td>Knowledge</td>
<td>127s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Multiple</td>
<td>Select</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#10</td>
<td>BIO1.LS1.1</td>
<td>Knowledge</td>
<td>134s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>88%</td>
<td>63%</td>
</tr>
<tr>
<td>#14</td>
<td>BIO1.LS1.8</td>
<td>Knowledge</td>
<td>108s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>88%</td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cluster #1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#25</td>
<td>BIO1.LS3.2</td>
<td>Knowledge</td>
<td>84s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>88%</td>
<td>NA</td>
</tr>
<tr>
<td>#26</td>
<td>BIO1.LS3.2</td>
<td>Knowledge</td>
<td>124s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>#27</td>
<td>BIO1.LS1.4</td>
<td>Knowledge</td>
<td>172s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>86%</td>
<td>29%</td>
</tr>
<tr>
<td>#28</td>
<td>BIO1.LS1.4</td>
<td>Knowledge</td>
<td>112s</td>
<td>88%</td>
<td>100%</td>
<td>100%</td>
<td>67%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cluster #2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#29</td>
<td>BIO1.LS3.3</td>
<td>Knowledge</td>
<td>139s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>75%</td>
<td>100%</td>
</tr>
<tr>
<td>#30</td>
<td>BIO1.LS3.3</td>
<td>Knowledge</td>
<td>106s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>83%</td>
<td>100%</td>
</tr>
<tr>
<td>#31</td>
<td>BIO1.LS1.5</td>
<td>Knowledge</td>
<td>134s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>13%</td>
<td>100%</td>
</tr>
<tr>
<td>#32</td>
<td>BIO1.LS1.5</td>
<td>Knowledge</td>
<td>169s</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>29%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>139s</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
<td>75%</td>
<td>83%</td>
</tr>
</tbody>
</table>

*Note.* Pass = Passage; Ques = Question; Ans = Answer; PK = Prior Knowledge; RTV = Return to Visual.

**Statistical Results**

One-tailed Pearson r correlations were conducted to determine the relationships between total raw scores on TNReady science items and median Maze scores for each grade. Results indicated that no significant relationships existed between these two measures for fifth grade students \( r(13) = -0.15, p > .05 \); eighth grade students \( r(12) = 0.23, p > .05 \); or students enrolled in high school \( r(8) = 0.12, p > .05 \).
One-tailed, independent sample T-tests were conducted to compare mean word counts and Flesch-Kincaid grade reading levels of testing items. Mean word counts between fifth grade items ($M = 193, SD = 81$) and eighth grade items ($M = 160, SD = 67$) did not differ significantly $t(21) = 1.07, p > .05$. Mean Flesch-Kincaid grade reading levels of fifth grade items ($M = 5.46, SD = 1.02$) and eighth grade items ($M = 6.68, SD = 1.29$) were significantly different $t(21) = -2.53, p < .05$. The mean Flesch-Kincaid grade reading level for eighth grade items was higher than that of the fifth grade items. Mean word counts between eighth grade items ($M = 160, SD = 67$) and high school items ($M = 147, SD = 67$) did not differ significantly $t(21) = 0.47, p > .05$. Mean Flesch-Kincaid grade reading levels of eighth grade items ($M = 6.68, SD = 1.29$) and high school items ($M = 10.24, SD = 1.76$) were significantly different $t(21) = -5.48, p < .05$. The mean Flesch-Kincaid grade reading level for high school items was higher than that of the eighth grade items.

The percentage of disagreements per item was calculated by dividing the total number of disagreements by the number of students in each grade. The average percentages, the minimum and maximum numbers, and the ranges of total disagreements by grade are presented in Table 8. The average percentages, the minimum and maximum numbers, and the ranges of total disagreements were similar across grades.
### Table 8

*Average Percentage of Disagreements per Item and Ranges of Disagreements*

<table>
<thead>
<tr>
<th></th>
<th>Average</th>
<th>Min</th>
<th>Max</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fifth Grade</strong></td>
<td>0.47</td>
<td>2.00</td>
<td>9.00</td>
<td>7.00</td>
</tr>
<tr>
<td><strong>Eighth Grade</strong></td>
<td>0.45</td>
<td>1.00</td>
<td>8.00</td>
<td>7.00</td>
</tr>
<tr>
<td><strong>High School</strong></td>
<td>0.58</td>
<td>1.00</td>
<td>7.00</td>
<td>6.00</td>
</tr>
</tbody>
</table>

### Post-Test Survey Results

Students in all three grade levels were administered the same five questions orally at the completion of testing items. The questions were designed to assist proctors in gathering information in regard to students’ experiences with the tests. Students were asked to indicate their general thoughts about the tests and whether or not the tests seemed like good measures of their science knowledge. Proctors also asked students whether they found anything confusing on the tests, whether they took computer-based tests often, and whether they had preferences between computer-based and paper and pencil tests.

**Fifth Grade**

Out of 13 respondents, two indicated that they found the test “hard” or “difficult” and one responded that it was “easy.” Eight of the respondents indicated the test was somewhere in between the two extremes (e.g., okay, good). One respondent mentioned the test contained some confusing words, and one respondent noted that they learned new material while taking the test. Most of the respondents indicated that the test seemed like a good measure of their knowledge in science. One respondent indicated that they felt the test was a good measure, although they had difficulty recalling previously learned material.
Asked whether anything was confusing about the test, seven respondents referred to specific questions that they found difficult. Two respondents indicated that some of the words or questions were confusing. The remaining four respondents did not report that anything was confusing for them. In response to the frequency of taking computer-based tests, most respondents indicated that they do not take computer-based tests very often. Some mentioned specific assessments they took on the computer (e.g., benchmark assessments, science, math, and reading). For some respondents, this was the second time they had taken a computer-based test, and one indicated that they had only begun taking computer-based tests during this particular school year.

Three respondents indicated that they preferred computer-based tests. One respondent reported that computer-based tests help them to be more organized, and another reported that tests are “easier” on the computer. Ten respondents indicated that they preferred paper and pencil tests. Reasons for this preference included unfamiliarity with test features (e.g., answer eliminators), difficulty with scrolling and referring back to visuals.

**Eighth Grade**

Out of 12 respondents, seven indicated that they found the test “easy” and three indicated that it was somewhere in between the two extremes (e.g., “it was challenging, but not challenging where I couldn’t do it at all”). One respondent indicated that it was “okay” and one respondent reported that it was “a normal test.”

All of the respondents indicated that the test seemed like a good test of their science knowledge. Seven respondents indicated that they did not find anything confusing on the test, and three referred to specific questions that they found confusing. One respondent mentioned that they found multiple select questions difficult, and another reported concerns with time
limits. In response to the frequency of taking computer-based tests, six respondents indicated that they do not take computer-based tests very often. Two respondents mentioned taking benchmark assessments on the computer, and four respondents reported that they take “some” tests on the computer (e.g., science, ELA).

Four respondents indicated that they preferred computer-based tests. Reasons for this preference included ease of clicking as opposed to completing bubbles or writing, being less distracted, and being less visually overwhelmed with seeing only one question at a time. Seven respondents indicated that they preferred paper and pencil tests. Reasons for this preference included feeling better able to concentrate, feeling safer (e.g., “something could go wrong” with computer-based tests) and preferring to write on the test paper itself. One respondent mentioned that taking tests in the same environment in which they learned the material helps with recalling information, as opposed to taking tests in an unfamiliar computer lab. Another respondent mentioned that “it depends on the subject.” They preferred paper and pencil for ELA tests and thought that computer-based tests are “good for science, social studies, and maybe math.”

**High School Students**

Out of eight respondents, seven indicated that they found the test somewhere in between the two extremes. In that category, three respondents mentioned difficult vocabulary, three others mentioned difficulty understanding the questions, and one respondent indicated that it was “okay.” Another respondent reported that the test is “always very different” in reference to content and level of difficulty. This respondent also indicated that some of the graphs were not helpful or useful.

Four of the respondents indicated that the test seemed like a good test of their biology knowledge. Two reported that it was not a good test of their biology knowledge due to “a lot of
inferencing” and the test containing some material not learned in class. Two respondents indicated that they found the test somewhere in between (e.g., “for the most part” and “it was a fairly good test”). Both of those respondents indicated that there were questions about material that they had not learned, and that they were taught the background knowledge to be able to answer the questions.

Three respondents mentioned the same question that they found difficult. Two respondents reported confusing wording of both questions and answer choices, and one respondent reported that the graphs were confusing and required more explanation. One respondent indicated that they did not find anything confusing on the test.

In response to the frequency of taking computer-based tests, all of the respondents reported taking computer-based tests quite frequently, either in specific subjects (e.g., math, ELA, science) or benchmark assessments. Two respondents indicated that they preferred computer-based tests. Reasons for this preference included ease of typing responses, use of tools (e.g. bookmark), and finding questions “easier and faster.” Two respondents reported their preference for paper and pencil tests. Reasons for this preference included feeling better able to concentrate, preferring to write on the test paper itself, and “remembering it better by writing because in class you didn’t type it.” Four respondents indicated that it depends on the subject. Out of those four respondents, one indicated that they prefer typing essays for ELA tests, and another respondent indicated that they prefer physically writing essays for the same subject.

**Discussion**

The goal of the current study was to help determine whether TNReady science test items evoked the intended cognitive processes and whether or not participants interacted with the instruments as the test developer anticipated. Another goal was to determine whether valid
inferences could be drawn from students’ scores on the tests as measures of their knowledge in science. The aim of the first research question was to determine whether the validity evidence based on cognitive processes indicated that the Tennessee state science assessments were technically sound measures of student knowledge of science according to the state science standards. The aim of the second research question was to determine the ways in which the questions evoked the intended cognitive processes.

The validity evidence based on cognitive processes gathered during the think aloud labs demonstrated that the Tennessee state science assessments were technically sound measures of student knowledge of science. Elements of Universal Design were embedded in the testing platform, including tools such as line highlighters, answer eliminators, and bookmarks (Thompson et al., 2002). Descriptive statistics revealed that students in all grades utilized the available enhancements. In addition, students applied strategies such as using prior knowledge and referring to visuals, despite disagreement among coders about specific strategy use. Results of the post-test survey revealed that the majority of students across all three grades reported that the tests were good measures of their science knowledge. Most students also indicated that the tests were in between the two extremes of difficult or easy. These results suggest that the students were interacting with the tests in ways that the test developer anticipated, and that the questions evoked the intended cognitive processes. TNReady science items are a direct assessment of science knowledge in accordance with grade level standards. This suggests that reasonable inferences can be drawn from students’ test scores as measures of their knowledge in science.

The aim of the third research question was to determine whether a relationship existed between Maze scores and number of correct responses to science items. There were no
significant relationships between the raw scores and median Maze scores across grade levels. In addition, mean word counts between fifth grade items and eighth grade items, and mean word counts between eighth grade items and biology items did not differ significantly. The Flesch-Kincaid grade reading levels of testing items were significantly different between fifth and eighth grade, and between eighth and biology. These findings suggest that the reading loads of the tests were not barriers for the students. It is reasonable to conclude that the tests do not require advanced reading abilities and the reading loads were appropriate across grade levels. These findings also demonstrate that the tests were successful in accomplishing the stated objectives, including measuring students’ knowledge of science according to state learning standards.

These results demonstrate that the TNReady science tests were developed appropriately and in accordance with Cognitive Load Theory (Beddow et al., 2011). The goals of test development under this theory include maximizing intrinsic load (cognition necessary to complete a task), using germane load in a purposeful manner (cognition that helps transfer input to long term memory), and minimizing extraneous load (ancillary requisite cognitive processes irrelevant to the task at hand). These concepts can be illustrated using a content standard and an item used on a past assessment. According to the Tennessee State Board of Education (2017), one of the fifth-grade content standards involves heredity, inheritance and variation of traits. Standard 5.LS3.1 states that students should be able to: “distinguish between inherited characteristics and those characteristics that result from a direct interaction with the environment. Apply this concept by giving examples of characteristics of living organisms that are influenced by both inheritance and the environment” (p. 43). The item asked students to identify which traits a mouse could inherit from its parents. The question included a brief, one-sentence question
stem, and four answer choices arranged in a vertical list. The intrinsic cognition necessary to answer this type of question includes the conceptual knowledge of inherited characteristics and those characteristics that result from a direct interaction with the environment. Extraneous load was minimized in this example in several ways. The question stem was short, and the list of traits (the answer choices) were simple phrases, thus reducing any unnecessary reading. The answer choice list was arranged in an organized manner, thus reducing unnecessary visual input. The majority of items on the TNReady science tests were similar to this example, confirming that the tests evoke the intended cognitive processes, measured the target constructs, and are technically sound.

**Prompts and Enhancements**

Students in fifth grade required more prompts than students in eighth grade or in high school. Students in fifth grade may have required additional time and instructions to understand the task demands during the think aloud labs. Students in eighth grade and students in high school may have understood the tasks more quickly and with fewer prompts than students in fifth grade.

High school students also utilized enhancements more frequently than fifth and eighth grade students. High school students may have more familiarity and experience with computer-based tests as opposed to paper-and-pencil tests. As a result, high school students may be able to navigate the computer-based tests and utilize the enhancements more efficiently than fifth and eighth grade students. High school students were more likely to utilize enhancements on items with cover sheets. Cover sheets often consisted of passages and images that required students to read and examine closely in order to answer the questions. As indicated on the post-test survey, two respondents stated that some questions required them to apply their science knowledge to
novel situations. High school students’ comfort with computer-based testing may have allowed
them to utilize enhancements whenever they encountered challenging questions.

**Raw Scores and Maze Scores**

The mean raw scores were similar across grades, and the scores tended to be slightly
higher among fifth grade students than eighth grade students and high school students. This may
be due to more difficult content and advanced curricula for students in higher grades. The
median Maze scores were the highest among high school students and lowest among fifth grade
students. It is possible that high school students are more experienced readers and thus able to
complete the task with more accuracy than fifth grade students.

Mean word counts between fifth grade items and eighth grade items, and mean word
counts between eighth grade items and biology items did not differ significantly, suggesting that
the average number of words per item were similar across grades. The Flesch-Kincaid grade
reading levels of testing items were significantly different between fifth and eighth grade, and
between eighth grade and biology. This result would be expected and would indicate that the
tests were developed appropriately, as the reading levels would increase as students moved into
higher grades.

**Student Strategy Use**

Students across grades were more likely to return to visuals on items that the test vendor
explicitly named this as a critical strategy. This finding demonstrates that the test-taking
population interacted with the tests the way the developer intended, and the appropriate cognitive
processes were being evoked. Students in fifth grade were more likely to reference a visual
(74%) than eighth grade students (59%). High school students were most likely to reference a
visual (83%) on items involving at least one visual. During the post-test survey, high school
students noted that the test required them to utilize prior knowledge to solve novel problems. The visuals often contained vital information necessary to solving the problems (e.g., charts). It could be that high school students referenced the visuals to help themselves answer the questions. This provides further evidence that the students were interacting with the test in the intended fashion.

Students across grades also utilized prior knowledge to answer most of the items. Overall, students in higher grades were more likely to report using prior knowledge (fifth grade: 50%, eighth grade: 65%, high school: 75%). Students in lower grades were more likely to take more time on average to complete items than students in higher grades (fifth grade: about two to four minutes, eighth grade: about one to two and a half minutes, high school: about one and a half to three minutes). These findings may be explained due to the cognitive demands of thinking aloud. It should be noted that this sample of fifth grade students performed worse on the Maze measure of reading fluency and comprehension, both overall and relative to benchmarks, compared with the eighth grade and high school students. It is possible that fifth grade students in this sample may have experienced more difficulty thinking aloud than eighth grade or high school students. In addition, during the post-test survey fifth grade students noted that they had little experience with computer-based tests. It could be that fifth grade students have less experience with testing than eighth grade and high school students, both overall and with computer-based tests. This could be a reason that students in eighth grade and high school students were more likely to utilize prior knowledge than fifth grade students, and a reason that fifth grade students took more time on average to complete items.

Transcript Coding

Both raters detected strategies that the students used including: using prior knowledge; using visualization or common sense and NOT referring to visuals; referring to visuals and NOT
using prior knowledge, visualization, or common sense; using prior knowledge, visualization or common sense AND returning to the visual. This finding demonstrates that students were using strategies and interacting with the instrument the way the test developer intended. Despite training and multiple practice opportunities, there were high percentages of disagreements between the raters on specific strategies used across all three grades. The strategies could have been interpreted in subjective ways from the transcripts. For example, one rater may have coded a behavior as a reference to a visual, and the other rater may have coded the same behavior as a reference to prior knowledge. In future studies involving think aloud labs, interrater reliability might be increased by further operationalizing student behaviors and providing additional examples of those behaviors during training.

**Post-Test Survey**

Asked their general thoughts about the test, the majority of students across grades felt that the level of difficulty was in between “hard” and “easy,” demonstrating that the tests were not excessively difficult. More students in eighth grade reported that they found the tests “easy.” Most students reported that the tests were good measures of their science knowledge. High school students were more likely to report that the tests required inferencing and critical use of background knowledge. Many students across grades mentioned the same questions and wording of answer choices that they found confusing. High school students also mentioned some of the graphs were unhelpful or required more explanation. Students in higher grades reported that they take computer-based tests more often than students in lower grades. More students in lower grades preferred paper and pencil tests than computer-based tests. Many high school students and one eighth grade student stated that it depends on the subject.
Limitations

Sample size was limited for this study and therefore may impact generalizability of the findings. A small sample size was necessary due to labor intensive nature of think aloud labs and transcript coding. The sample was limited on some demographic variables, although it was representative of the target population of test takers and the overall demographics of both school districts. Most of the students were White and few qualified for free and reduced priced lunch. The sample also included high performing students that were native English speakers due to the tasks required during think aloud labs. It is difficult to determine the ways in which students with disabilities and English language learners would interact with the tests. Additionally, the roles that accommodations would play for students with disabilities are unclear. For example, none of the students that took part in the think aloud labs had other accommodations such as breaks, or having the questions or directions read and reread aloud by a proctor.

Implications for Practice

Thurlow, Lazarus, Albus, and Hodgson (2010) reported that computer-based testing is becoming more prevalent throughout the United States. According to Kettler et al. (2011), students should be provided with opportunities to practice modules for computer-based tests to ensure equitable access. Many students in the lower grades reported that they do not take computer-based tests often. It would be beneficial to students in all grades to have more opportunities to practice computer-based testing. Short practice sessions could take place at the beginning of computer-based tests such as benchmark assessments or be embedded into instructional time.

Gorin (2006) emphasized the importance of clearly defined constructs of interest throughout the test development process. Kettlerin-Geller (2008) opined that teachers could
determine the appropriate accommodations on an item-by-item basis, depending on student need and the construct of interest. According to the results of the post-test survey, many students reported that their preferences for computer-based versus paper and pencil tests depends on the subject. Teachers and school districts could consider giving students flexible options for the mode of testing depending on student preferences, needs, and subjects.

According to post-test survey responses, students reported that certain visuals or graphs were unhelpful or needed more explanation. Test developers should be strategic and select graphics that enhance test takers’ understanding so they can properly demonstrate their mastery of the target constructs. Additionally, students across grade levels mentioned the same questions and wording that they found confusing. These items should be examined and reworded according to principles of Universal Design, with particular attention to “maximum readability and comprehensibility” (Thompson et al., 2002, p. 6).

Conclusions

Overall, the findings from this study provided sufficient validity evidence based on response processes to support the conclusions that the TNReady science test items evoked the intended cognitive processes, participants interacted with the tests as the developers intended, and the assessments are technically sound measures of students’ knowledge of science. Reading load was not a barrier, as no statistically significant relationships existed between the median Maze scores and the number of correct responses to science items. The average number of words per item was similar across grades. The Flesch-Kincaid grade reading levels of testing items were significantly different between fifth and eighth grade, and between eighth and high school, indicating that the reading levels of the tests were commensurate with students’ grade levels.
Students in lower grades needed more prompts during the think aloud labs and performed worse on the Maze measures. Students in higher grades utilized more enhancements (e.g., line readers, highlighters) and referenced prior knowledge more often. Raw scores were higher in lower grades, perhaps reflecting an increase in difficulty of content in higher grades. Students were more likely to reference visuals whenever the test vendor explicitly named this as a critical strategy. Proctors detected student strategy use on all items despite high levels of disagreement on specific strategy use. Survey results indicated that most students found the tests to be in between “hard” and “easy,” illustrating that the tests were not excessively difficult. Students reported that they felt the tests were good measures of their knowledge of science. Based on survey results, test developers should ensure that the wording of each item is clear, and the graphics are helpful and necessary. Schools should consider offering students more practice with computer-based tests and flexible options for test delivery, as is feasible. Future studies could include a more diverse sample in terms of race and ethnicity, socioeconomic status, English Language Learners, and students with disabilities.
References


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*


input in developing alternate assessments based on modified academic achievement 

Retrieved from https://www.tn.gov/sbe/committees-and-initiatives/standards-
review/science.html.

scale assessments (Synthesis Report 44). Minneapolis, MN: National Center on 
Educational Outcomes.

Practices and considerations (Synthesis Report 78). Minneapolis, MN: University of 
Minnesota, National Center on Educational Outcomes.


factors that affect assessment validity for diverse learners: Results from a large-scale 
cognitive lab. Learning and Individual Differences, 16, 267-276.