Spring 2015

Statement verification for science : examining technical adequacy of alternate forms for screening decisions

Jeremy W. Ford
University of Iowa

STATEMENT VERIFICATION FOR SCIENCE: EXAMINING TECHNICAL ADEQUACY OF ALTERNATE FORMS FOR SCREENING DECISIONS

by

Jeremy W. Ford

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Teaching and
Learning (Special Education)
in the Graduate College of
The University of Iowa

May 2015

Thesis Supervisor: Professor John L. Hosp

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Jeremy W. Ford

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Teaching and Learning (Special Education) at the May 2015 graduation.

Thesis Committee:

John L. Hosp, Thesis Supervisor

Allison L. Bruhn

Kristen N. Missall

William J. Therrien

Suzanne Woods-Groves

To Rachel, Isaiah, Caleb, and Owen

Just keep swimming,
Just keep swimming,
Just keep swimming, swimming, swimming …

Dory
Finding Nemo

ACKNOWLEDGMENTS

The struggle to “just keep swimming” over the course of my doctoral studies would not have been successful without the encouragement and support of many. Much of this support came from individuals in the special education program here at The University of Iowa. I would like to especially acknowledge my advisor, Dr. John Hosp, for his mentorship over the last four years and his guidance regarding my dissertation. I would also like to thank my committee, Dr. Allison Bruhn, Dr. Kristen Missall, Dr. William Therrien, and Dr. Suzanne Woods-Groves for their time and direction related to this study and other experiences related to my training. Dr. Youjia Hua also deserves recognition for his willingness to involve me in his research early in my doctoral program. Many thanks are also necessary for those who I’ve studied with over the last four years. I truly believe the members of one’s cohort are often the only individuals capable of knowing what you are going through. I appreciate your help getting out of several trees and hope I’ve been assistance to you as well.

Last, I want to acknowledge my family. To my children who experienced this journey with me – Isaiah, Caleb, and Owen – I hope what you recall of this time in your life is black and gold and Tigerhawks! To Rachel, no one deserves more acknowledgement than you. From immediately telling me to apply to my doctoral program when the opportunity presented itself to your support through my dissertation, I simply cannot thank you enough for all that you mean in my life. I’ll love you forever, and ever, and always, and a day.

ABSTRACT

The *Rising Above the Gathering Storm* report (National Academy of Sciences, 2007) emphasizes a need for improved science education in United States schools. Instruction, informed by assessment, has been repeatedly demonstrated to be effective for increasing students' performance. In particular, the use of curriculum-based measurement (CBM) to assist with making screening decisions has been shown to increase the likeliness of students meeting meaningful outcomes. While CBM tools for assisting with making screening decisions in reading, mathematics, and written language have been well examined, tools for use in content areas (e.g., science and social studies) remain in the beginning stages of research. In this study, two alternate forms of a new CBM tool (Statement Verification for Science; SV-S) for assisting with making screening decisions regarding students' science content knowledge, is examined for technical adequacy.

A total of 1,545 students across Grades 7 (N = 799) and 8 (N = 746) completed Forms A and B of SV-S the week prior to, and within two weeks after, a statewide high-stakes test of accountability including Science, Reading, and Mathematics. Obtained data were used in order to examine internal consistency and test-retest with alternate forms reliability as well as evidence of criterion- and construct-related validity. Promising results were found for reliability, in particular internal consistency, while results related to evidence of criterion- and construct-related validity were less than desired. Such results, along with additional exploratory analyses, provide support for future research of SV-S as a CBM tool to assist teachers and other educators with making screening decisions.

PUBLIC ABSTRACT

The fields of Science, Technology, Engineering, and Mathematics (STEM) play an increasing role in the everyday lives of those living in the United States and around the world. Yet, for students in the United States, evidence suggests knowledge related to STEM fields is not keeping pace with such increased influence. As a result, teachers and other educators are faced with the challenge of identifying students likely, and not likely, to meet future outcomes related to science and related disciplines. These types of decisions have been called screening decisions and their purpose is to identify which students require additional instructional support in order to meet expectations (e.g., learn grade level science content by the end of the school year).

Curriculum-Based Measurement (CBM) is a measurement technology which includes many tools which can be used to assist teachers and other educators with making screening decisions in reading, mathematics, and written language. However, in content areas (e.g., science and social studies) tools are still being developed.

This study provides potential evidence that a new CBM tool for content areas, Statement Verification for Science (SV-S), possesses the necessary characteristics to assist with making screening decisions. A total of 1,545 students across Grades 7 and 8 participated in this study. In this study, students' performance on two alternate forms of SV-S and a statewide high-stakes test of accountability were examined.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Importance of Science Education

A call for improved science education in public schools in the United States has been made. The National Academy of Sciences (2007) report, *Rising Above the Gathering Storm,* emphasizes a need for improvement by stating "virtually all quality jobs in the global economy will require certain mathematical and scientific skills" (p. 135). The recent emphasis in the Science, Technology, Engineering, and Mathematics (STEM) fields in recent years reflects a call for investing in the scientific literacy of students. Such an investment seems warranted given the changes in today's society due to the rapid rate of scientific and technological discovery. In order for individuals to participate in society and benefit from increasing scientific and technological discoveries, however, a basic level of scientific literacy will be necessary.

One could argue in a changing society, fueled by rapid scientific and technological development, what is considered to be a basic level of scientific literacy will need to rise. If such a premise is accepted, the performance of U.S. fourth and eighth graders on the Trends in International Mathematics and Science Study (TIMSS) may be cause for alarm. That is, no measureable difference in the average science score obtained by U.S. fourth graders from 1995 (542) to 2011 (544) on TIMSS has been observed, nor has any measureable difference in the average science score obtained by U.S. eighth graders from 1995 (513) and 2011 (525). With recent educational initiatives focused on improving students' reading and mathematics performance, as measured by similar tests, it may well be time to call for initiatives to increase students' knowledge of science as measured by such tests as well.

One may also argue that the results of international testing have been overstated and used for political reasons (Salzman & Lowell, 2007). This may be the case as the average science score for U.S. fourth graders of 544 in 2011 was higher than the international TIMSS scale average of 500. Further, U.S. fourth graders ranked in the top 10 internationally with 6 education systems ranked ahead and 3 not being measurably different. In addition, U.S. fourth graders ranked higher than 47 other education systems. Results for U.S. eighth graders are similar. Their average science score of 525 was higher than the TIMMS scale average of 500, they ranked in the top 23 education systems with 12 having higher averages and 10 being not measurably different, and they scored higher than 33 education systems.

Regardless of whether or not one focuses on the static performance of students in the U.S. as an indication of needing sweeping changes to science education or if one highlights U.S. students' relative higher performance compared to the international community as a reason to make any changes with caution, a closer look at the TIMSS results does point to inequities in the U.S. educational system in need of being addressed. That is, a great deal of variability exists within the U.S. regarding student performance on TIMSS. For example, in eighth grade eight U.S. states were involved with TIMSS as benchmarking participants (Alabama, California, Colorado, Florida, Indiana, Massachusetts, Minnesota, and North Carolina). Three states (Colorado, Massachusetts, and Minnesota) were among the 12 education systems obtaining a higher average science score than the eighth graders in the U.S.

Further, variation within the U.S. has been observed when considering poverty. Using TIMSS results from 2007, the National Center of Education Statistics (2009) reported U.S. fourth graders who attended high poverty schools, where at least 75% of students are eligible for free or reduced-price lunch, obtain average science scores on the TIMSS of 477, less than the

scale average of 500. Further, U.S. fourth graders who attended the lowest poverty public schools, where less than 10% of students are eligible for free or reduced-price lunch, obtained average science scores of 590. The average U.S. fourth grade science score for students attending low poverty public schools was higher than the U.S. average of 539 as well as higher than the average score from top ranked Singapore (587).

Though somewhat more encouraging, the statistics for U.S. eighth graders are similar. U.S. eighth graders who attended public schools where less than 50% of students were eligible for free or reduced-price lunch scored higher than the TIMSS scale average of 500. However, U.S. eighth graders who attended high poverty schools obtained an average science score of 466. Further, the average U.S. eighth grade science score for students attending low poverty public schools was 572, higher than the U.S. average of 520 as well as the average score from top ranking Singapore (567).

Improving Science Education

The *Rising Above the Storm* report suggests improving instructional practices regarding science content will result in a scientifically literate citizenry and will encourage individuals to pursue science- and mathematical-based careers. Given changes in society due to scientific and technological findings, the focus of STEM initiatives in public schools makes sense. That is, seeing that students obtain the skills needed for participating in society is oft provided as a function of public schooling (Curtis, 1991).

In order to maximize its effectiveness, instruction must be informed by student need. Determining the instructional needs of students requires the use of reliable and valid methods of assessment, a systemic process for collecting information. As assessment refers to a process for collecting information, it is important to note several methods are available to accomplish such a

purpose. Common methods for collecting information in school settings include: (a) Reviewing student work, (b) Interviewing students and teachers, (c) Observing students in a classroom setting, and (d) the use of Testing. Collectively, these methods are often referred to as RIOT procedures (cf. Hosp, Hosp, Howell, & Allison, 2014).

Since the late 20th century a focus on assessment procedures for accountability purposes has developed. Though focusing on assessment procedures for accountability has legal and ethical implications (Saliva, Ysseldyke, & Bolt, 2009), it has little to do with informing teachers on how to best instruct students. However, the use of assessment procedures has importance beyond accountability purposes. For example, the use of assessment procedures for formative decision making is important for increasing students' learning. Formative decision making occurs when assessment procedures are used to collect information about students' learning that provides feedback to students and teachers during instruction rather than at its conclusion (Bell & Baker, 1997).

Assessment for Formative Decision Making

Effective use of assessment procedures generally positively influences student learning (Crooks, 2002; Gipps & James, 1998) and this holds true for formative decision making as well. In their seminal review, Black and Wiliam (1998) concluded the research to "conclusively" demonstrate the use of assessment procedures for formative decision making to improve students' learning. Further, Black and Wiliam (1998) concluded the use of assessment procedures for formative decision making to be amongst the most effective educational interventions available to teachers. Hattie (1999), in a meta-analysis examining the relative effects of different teaching approaches, similarly found the most powerful moderator for enhancing student achievement to be feedback, a key component of formative decision making.

Responsible formative decision making, in particular for high-stakes decisions (e.g., those related to entitlement to special programs) must rely on tools possessing adequate technical adequacy. That is, tools with established evidence of validity and reliability are necessary in order to increase the likeliness of accurately determining students' response to instruction.

Characteristics of Technical Adequacy

Technical adequacy consists of two concepts, validity and reliability. The concept of validity is concerned with the meaningfulness (i.e., usefulness or appropriateness) of the specific inferences one can make based on the results of a tool (AERA, 1997). The concept of reliability is concerned with the extent to which results from a tool are free from error (Gronlund & Waugh, 2009). Both concepts are important to consider when developing, or using, a tool to measure student learning. However, given the possibility that an accurate (i.e., error free) tool could provide meaningless results, the concept of validity is of greater concern. The traditional view of the concept of validity held there to be several types thereof, but more recent conceptualizations describe validity as a unitary concept based on various forms of evidence (Messick, 1989). In addition, evidence of validity is not measured, but inferred. For example, if one wants to use a tool to predict, or estimate, performance on another tool ideally evidence would be available to infer the use of the tool for predicting exists.

Curriculum-Based Measurement (CBM) includes tools that are frequently used for formative decision making (Hosp, et al., 2014) and the technical adequacy of CBM is well established in the empirical literature (Foegen, Jiban, & Deno, 2007; McMaster & Espin, 2007; Wayman, Wallace, Wiley, Tichà, & Espin, 2007). In addition, the results from using CBM allows one to make low inference decisions regarding students' skills. The use of tools allowing for low inference instructional decision-making is preferred over tools which only allow for high

inference instructional decision-making as the former can be considered more closely aligned to the curriculum (Hosp et al., 2014; Hosp, Hosp, & Howell, 2007).

Curriculum-Based Measurement

CBM was developed during the late 1970s and early 1980s at the Institute for Research on Learning Disabilities (IRLD) at the University of Minnesota. The IRLD focused on developing and assessing the technical adequacy of indicators for measuring student performance in the curriculum for the purpose of improving instructional decision making. Several indicators were identified in reading, mathematics, written expression, and spelling as demonstrating acceptable levels of reliability and evidence of validity for decision making (Espin, McMaster, Rose, & Wayman, 2012).

Most of the research involving CBM has focused on reading with elementary aged students (Wayman et al., 2007). However, as students leave elementary school and enter middle and high school many experience difficulty – in particular students without strong reading skills. This heightened difficulty can be attributed to the change of emphasis of learning to read in elementary school to reading to learn in secondary school (Chall, 1999). Statistics regarding the reading proficiency of secondary students are alarming. The reading ability of secondary students, in particular their ability to read and comprehend content material such as science, is especially a concern given the recent emphasis from the Common Core and the Next Generation Science Standards in content area performance.

Secondary Students and Reading

Deshler, Palincsar, Biancarosa, and Nair (2007) stated there is a concern the U.S. is in the midst of an adolescent literacy crisis. That is, while students in secondary schools are expected to read and comprehend challenging text in content areas (Alvermann, 2002), many struggle with

the necessary prerequisite reading skills (i.e., decoding, fluency, vocabulary, comprehension) to do so (Kamil, 2003). More specifically, Biancarosa and Snow (2006) reported that approximately 70% of U.S. students in Grades 4–12 struggle to read on grade level. Given the importance of reading both in and out of schools (National Reading Panel Report, 2000) such statistics are certainly a concern.

Research into extending the application of CBM to reading in content areas (an emphasis of secondary schools) exists. This research has explored reading aloud using content area material, maze selection, and vocabulary matching (Beyers, Lembke, & Curs, 2013). Most of the research using CBM with secondary students has focused on vocabulary matching as Espin and Foegen (1996) found it to account for a significant amount of the variance for multiple tasks of comprehension when entered first into a regression model. In addition, Marcotte and Hintze (2009) found the Sentence Verification Technique to contribute significantly to an overall model of reading comprehension after controlling for reading aloud. While the SVT is not a CBM task, this finding may provide a direction for additional research regarding CBM in content areas.

However, despite the potential of vocabulary matching and the SVT, there are issues regarding their use. For example, the common metric used in vocabulary matching (correct responses) is not sensitive to change and thus may not be effective for informing instruction. Regarding the SVT, the primary issue may be the difficulty in accounting for the background knowledge a student brings to the task. Therefore, despite their promise, alternatives to these tools should be explored. One such tool is Statement Verification for Science (SV-S).

Purpose of the Study

The purpose of this study is to examine the technical adequacy of alternate forms for Statement Verification for Science (SV-S). In completing SV-S students silently read factual

statements regarding science content knowledge and indicate if the statement is correct or incorrect. Statements are related to the Iowa Common Core Science Standards and the Next Generation Science Standards. Students are given 3 minutes to complete the task. The design of this tool addresses the issues discussed regarding vocabulary matching and the SVT as a fluency score will be able to be calculated and background knowledge is only important as it applies to knowledge of content within science standards respectively.

CHAPTER TWO

LITERATURE REVIEW

Chapter Overview

Given the changes in today's society due to scientific and technological advances, and the challenges faced by many secondary students regarding reading, much research has focused on ways to improve student learning outcomes. This chapter will address this research starting with the role of assessment procedures, in general and specifically regarding those related to formative decision making, in increasing student achievement; the development and application of CBM as a technology for formative decision making; the potential role CBM could play in formative decision making regarding science education; various studies involving the use of CBM in content areas; and the rationale for exploring SV-S as a new tool for the purpose of formative decision making in science education.

The Role of Assessment in Student Learning

Information gathered using various assessment procedures can be used for evaluation (i.e., making educational or instructional decisions). Teachers use assessment procedures to inform instruction when considering screening, progress, diagnostic, and outcomes decisions. It can be helpful to think of these different decisions in relation to how they answer specific questions. Hosp (2011) suggested: (a) screening decisions focus on "Which students are not currently meeting expectations?", (b) progress decisions focus on "How much progress are students making toward meeting expectations?", (c) diagnostic decisions focus on "What content or leaning objective is of concern?", and (d) outcome decisions focus on "Have students met learning expectations based on the instruction they were provided?".

Screening, progress, diagnostic, and outcome decisions have both external (i.e., outside of the classroom) and internal (i.e., inside the classroom) purposes (Hosp, 2011). Although both purposes are important for teaching and learning, internal purposes have clearer implications for teachers. That is, assessment assists teachers with making screening decisions by identifying students unlikely to meet a future, meaningful outcome (e.g., Good, Simmons, & Kame'enui, 2001; McGlinchey & Hixson, 2004; Nese, Park, Alonzo, & Tindal, 2011) in order to determine which students should receive additional instruction. In addition, assessment assists teachers with making progress decisions by providing a means to determine how students are responding to instruction (e.g., Fuchs, Fuchs, Hamlett, & Stecker, 1991; Stecker & Fuchs, 2000). Further, assessment assists teachers with making diagnostic decisions by identifying the instructional focus needed for a student (e.g., Stecker, Fuchs, & Fuchs, 2005). Last, assessment assists teachers with making outcome decisions by determining whether students have obtained proficiency and/or mastery (e.g., Fuchs, Deno, & Mirkin, 1984; Tindal, 1992).

Assessment procedures can be used for both summative and formative decision making purposes. Torgesen and Miller (2008) referred to summative decisions being focused on "assessment *of* learning" and formative decisions being focused on "assessment *for* learning." Though the definitions provided by Torgesen and Miller (2008) are helpful for highlighting the primary difference between summative and formative decision making, in reality the distinction is not so clear. Hosp (2011) suggested that summative and formative decision making occur along a continuum and, depending on their nature, assessment procedures for making screening, progress, diagnostic, and outcome decisions can occur at varying points along this continuum. For example, universal screening decisions can be viewed as being more summative in nature given their less frequent occurrence as results can be used to estimate what was learned during

the time since the previous collection of universal screening data. Whereas screening decisions made as a result of considering more frequently administered unit pretests can be viewed as being more formative in nature given results could be used to design future instruction (e.g., re-teaching concepts many students are observed to not understand or quickly reviewing concepts most students are observed to understand). While noting rate of occurrence is useful for distinguishing between summative and formative decision making, it is important to highlight it is the purpose of the decision being made that defines whether summative or formative decision making is occurring. Thus, formative decision making can only be considered to occur if the information obtained is used to enhance students' learning (Black, 1993). That is, collecting pretest data is not an example of formative decision making, using the results of such data to positively affect students' learning needs to take place as well.

Although the use of both formative and summative decisions are important, the use of assessment procedures for formative decision making has been repeatedly shown to be an effective educational practice for improving student learning outcomes. For example, use of assessment procedures for the purpose of formative decision making, across a range of studies including many different areas of content, has been observed to yield significant achievement gains for individuals five years old to undergraduate university students (Black & Wiliam, 1998). Further, Hattie (2009) found the use of assessment procedures for formative decision making to inform instruction to have an effect size of .90 compared to instruction not informed by its' use.

CBM is a measurement technology, encompassing several tools, which has been demonstrated to have the characteristics needed to be used responsibility for formative decision making as it relates to students' learning (cf. Hosp et al., 2007).

Curriculum-Based Measurement

The origins of CBM can be found in the development of Data-Based Program Modification (DBPM; Deno & Mirkin, 1977). The use of an inductive approach for determining instructional needs of students was a major component of DBPM. That is, Deno and Mirkin (1977) proposed that determining what instructional intervention will work, especially for academically struggling students, should be considered a hypothesis to be empirically tested. A second essential component of DBPM was an emphasis on the use of time-series analyses for testing instructional hypotheses given their unique ability to gauge the effects of intervention. In addition, Deno and Mirkin (1977) also proposed that monitoring students' progress should focus on "vital signs" of educational development. The research to identify such "vital signs," spearheaded by the IRLD at the University of Minnesota during the late 1970s and early 1980s, led to development of CBM.

To adequately serve the "vital signs" function, Deno (2003) identified nine attributes of CBM. One attribute of CBM is its alignment with the curriculum. That is, CBM is founded on the idea that assessment and decisions regarding instruction are curriculum-referenced (i.e., CBM tools reflect the skills students are taught, Deno, 1985; Deno 1986). For example, a fourth grade CBM probe for math computation would include all the computation skills a student would be taught during fourth grade. As students progress through the school year, they are expected to perform better on equivalent probes as they are introduced to, and learn, new skills. Thus, CBM can be used to measure the skill level of students in the curriculum on which they receive instruction.

A second attribute of CBM is that it is technically adequate. This attribute is an important distinction from most tools used by classroom teachers which do not have established reliability

and validity. For example, a great deal of research was conducted in the 1980s and 1990s demonstrating the relation between the number of words read correctly (WRC) from reading aloud from connected text for one minute and multiple criterion measures of reading (Wayman, Wallace, Wiley, Tichà, & Espin, 2007). This metric is also commonly referred to as oral reading fluency (ORF). However, Hosp and Suchey (2014) note the term oral passage reading (OPR) better reflects the nature of the task of calculating the number of WRC from a one minute reading of connected text. That is, the definition of "fluency" includes an aspect of prosody which is absent in the task measured by WRC. Therefore, I will continue to use the term OPR where appropriate hereafter.

A third attribute of CBM is its common use for making criterion-referenced decisions. Criterion-referenced decisions, as opposed to norm-referenced decisions, are preferred for determining students' level of proficiency with an academic skill rather than determining how their current skills compare to peers. While CBM being used for criterion-related decision making is related to the first attribute of CBM (i.e., it being curriculum-referenced), the distinction to be made here is the focus of this attribute on using tools for decision making rather than on the nature/development of a particular tool. Using a criterion-referenced decision making framework is important when making screening decisions. That is, it is prudent to use an empirically derived cut score to determine if a student is likely to meet expectations on a future meaningful outcome rather than simply comparing that student's performance to their peers. This is prudent as a cut score in this scenario represents a level of proficiency that allows one to infer the student's skill level is appropriate for the given time. It is possible, when using a norm-referenced framework for making screening decisions, that a student could be comparable to

his/her peers (i.e., average), but not possess an adequate skill level if the norm group does not possess an appropriate skill level as well.

A fourth attribute of CBM is the use of standard procedures for administration and scoring (i.e., the tools used in CBM are standardized). Thus, if educators desire to share information regarding a student, CBM tools must be administered and scored according to prespecified criteria. In addition, while many CBM tools are available for purchase, standardized guidelines are also available for those interested in developing their own (cf. Shinn, 1989).

The use of performance sampling is a fifth attribute of CBM. Based on tasks which are operationally defined, CBM measures correct and incorrect student behaviors. Doing so allows for low inference decisions regarding the meaning of obtained results. Making a low inference decision about a student's skill level is possible with CBM as the tools being used directly measure the academic behavior of interest (i.e., the skills contained in the curriculum students are expected to learn during the school year). High inference decisions are necessary when using many tools used to measure student's academic skill level. This is because such tools require conjecture regarding academic behavior not observed, or observed infrequently, during data collection.

For example, results from a tool requiring the use of high inference decisions may provide a score related to a student's computation skills but the tool may only include a handful of computation items. Further, the handful of items included regarding computation likely include only one or two – if any – which address certain types of computation (e.g., two digit by two digit addition without carrying or two digit by one digit subtraction with borrowing). Such an inadequate sample of computation items makes generalizing to the student's true

computational skills, and identifying specific areas of concern (i.e., diagnostic decision making) difficult.

Conjecture is abated when using tools associated with CBM as the data obtained are specific to the skill of interest. That is, a score of 15 correct digits (CD) means that a student obtained 15 CD on a CBM math computation probe during the time the probe was administered (e.g., four minutes). Further, if an empirically derived cut score notes a score of 30 CD as being necessary to predict the student will be observed to be proficient on a future meaningful outcome, then little inference is needed to determine the student requires additional instruction in computation. In addition, procedures are available to use CBM math computation to make diagnostic decisions regarding specific skills instruction may need to focus on (Hosp et al., 2007).

A sixth attribute of CBM is the application of predetermined rules for decision-making. The use of empirically based rules for decision-making allow teachers and other educators to make instructional decisions based on student performance. Ardoin et al. (2012) note there are two types of decision rules that are commonly applied to OPR time series data to evaluate students' response to instruction. Such rules are applied to evaluate students' response to instruction in other academic areas as well (cf. Hosp et al, 2007). One common type of decision rule is the data point decision rule. Typically this involves comparing a students' observed rate of growth to an aim (or goal) line. The aim line represents the desired rate of growth from an individual as it connects his/her current skill level (i.e., their baseline performance data) and a fixed, future, datum point representing expected skill level (e.g., an end of the year benchmark) at a later time. A decision to change instruction occurs when three to five consecutive data points

are observed to be below the aim line. When three to five data points fall above the aim line then a decision to raise a goal may be made.

The second common type of decision rule applied to evaluating students' response to instruction involves the use of trend line analysis, of which several different methods (e.g., ordinary least squares regression, quarter-intersect, split-middle, and Tukey) are available (Ardoin et al., 2012). All methods of trend line analysis involve estimating a student's rate of growth, represented by determining the slope of a student's progress. The trend line is then compared to the aim line in order to assist educators with making decisions regarding students' response to instruction similar to the data point decision rule approach (Parker & Tindal, 1992).

The use of repeated measurement over time is a seventh attribute of CBM. A benefit of the use of repeated measurements over time is it allows educators to make decisions about the effectiveness of their instruction in relatively short periods of time. That is, within several weeks repeated measurement of student performance may indicate a need to modify instruction rather than such a realization occurring at the end of a semester or school year. In fact, CBM was initially developed to provide a technically adequate method for educators to monitor students' progress toward meaningful educational outcomes (Deno, 1985). Research has found that the use of CBM by teachers helps to accomplish greater achievement rates for their students (Stecker, Fuchs, & Fuchs, 2005; Fuchs, Deno, & Mirkin, 1984).

The eighth and ninth attributes of CBM both relate to its efficiency. One reason why CBM is efficient is the ease with which individuals can be trained to administer and score tools. Further, a small amount of time (e.g., one minute for OPR) is needed to administer many CBM tools and calculating student performance is done quickly using simple math computation. Using assessment tools that are efficient is ideal as it allows for more time for instruction rather than

assessment (Hosp & Ardoin, 2008). The final attribute of CBM is that student results can be summarized efficiently. Often this includes the use bar and line graphs. Such simple visual representations of student performance can be helpful for relaying a student's skill level to other educators, parents, or even the student.

As mentioned previously, CBM was developed during the late 1970s and early 1980s at the University of Minnesota's IRLD to be used for monitoring students' growth in academic areas as well as evaluating the effects of instructional programs (Deno, 1985). CBM was also developed as a key component for problem-solving educational models such as DBPM and its progeny (e.g., Response to Intervention, Multi-Tiered Systems of Support). It was intended that individuals using such a model, when making decisions regarding a student's eligibility for special education, consider observed academic difficulties as a problem to be solved rather than assign the cause of such difficulty to intrinsic, unchallengeable characteristics of the student (Deno, 1990).

The impetus for the initial development of CBM was twofold (Deno, 1985). First, was the aforementioned need for tools with the necessary characteristics of technical adequacy (i.e., reliability and evidence of validity) for use in making instructional decisions. Second, was the need for practicality. That is, because the tools associated with CBM were to be used for teachers to frequently measure the effects of instructional programs they needed to be simple, efficient, interpreted easily, and inexpensive. Thus, CBM tools were designed to be short samples of student produced work which were valid and reliable regarding the broader academic domain they were associated with (Wayman et al., 2007). The research on CBM in reading (Wayman et al., 2007), mathematics (Foegen, Jiban, & Deno, 2007), and written language (McMaster & Espin, 2007) supports the claim that these needs have been addressed.

Wayman et al. (2007) note OPR, maze selection, and word identification to be the CBM tools most commonly used in reading. As described above, OPR is measured using the WRC metric. When completing a traditional maze selection probe (Hosp, et al., 2007), students silently read a passage for 1-3 minutes. The first and last sentences of the passage remain intact while approximately every 7th word is deleted in the passage. Deleted words are replaced with a choice of three replacement words, the answer and two distractors. Students circle the word they believe best fits the sentence. The total number of correct restorations is typically how performance on maze selection is measured (Silberglitt et al., 2006). Word identification involves students reading from a list of high-frequency words, often for one minute, with the number of WRC counted (Deno, Mirkin, & Chiang, 1982).

Regarding the research in CBM and mathematics, Foegen et al. (2007) noted disagreement on the best approach to developing tools (i.e., using the curriculum sampling or robust indicators approach). However, they note the largest number of studies has been conducted with elementary aged student using both approaches. They further noted that both approaches have been used with secondary students while the robust indicator approach has been most studied in early mathematics studies. CBM in early mathematics includes tools related to early numeracy. Clarke and Shinn (2004) described some of these tools. For example, for Missing Numbers students are presented with a box (on a sheet of paper) which contains three numbers and a blank. The numbers consist of a pattern (e.g., counting by 1's or 5's) and students are to identify the missing number from the pattern. The number correct a student obtains in a minute is recorded. For students in elementary school and beyond Math Computation is a common CBM tool for measuring student's mathematics skills. Math Computation is conducted by determining the number of CD a student obtains when answering computational problems for

a specific period of time (Hosp et al., 2007). Concepts and Applications include skills related to measurement, time, and graph interpretations and have been developed to use for students in upper grades where the focus of the mathematics curriculum is more than computation (Hosp et al., 2007).

CBM for Written Language involves students responding in writing to an instructional level story starter for 3 minutes and being scored on several simple indices of performance (Hosp et al., 2007). Common approaches to scoring include counting the total number of words written (TWW), counting the total number of words spelled correctly (WSC), and counting the total number of correct writing sequences (CWS). These first two indices are self-explanatory while CWS is defined as "two adjacent, correctly spelled words that are acceptable within the context of the [written] phrase to a native speaker of the English Language" (Videen et al., 1982). McMaster and Espin (2007) examined evidence of reliability and validity for CBM for Written Language. One finding worth highlighting from their review is that while elementary students at the group level made significant gains over the course of the school year in terms of the TWW, CWS, and CWS some students improved due to quantity in writing, not quality (and vice versa). As a result, McMaster and Espin (2007) identified a need for a multifaceted approach to measuring elementary students' written language skills for the purpose of educational decision making. A similar finding, in regard to secondary students, concluded that simple, countable indices (i.e., TWW, WSC) are not sufficient indicators of writing skill. As a result, additional research has focused on investigating more complex metrics (e.g., correct minus incorrect writing sequences).

Therefore, given the benefits of using assessment procedures for formative decision making for improving student outcomes, the ability of the tools associated with CBM to evaluate

students' response to instruction, and the need to increase the scientific literacy of U.S. students, it is valuable to explore potential uses of CBM in the content area of science.

Curriculum-Based Measurement in Content Areas

Research on using CBM in content areas has been conducted in science and social studies, focusing on: reading aloud from content area material, vocabulary matching, and maze selection. Another approach, the sentence verification technique (SVT), has been examined as a means for measuring students' reading comprehension. SVT may possess characteristics that, once modified, could measure students' content knowledge using CBM principles. Pros are present for using each of these approaches and will be discussed below. However, limitations are also present. These limitations will also be presented and used to frame the rationale for a new approach to CBM in content areas.

Reading Aloud

The initial research into using CBM in content areas included examining the validity of reading aloud (i.e., OPR) using biology text (Espin & Deno, 1993). More specifically, Espin and Deno (1993) examined the relation between reading aloud (from both English and science texts) and a researcher developed classroom study task. Results demonstrated low moderate correlations for both English (r = .37) and science (r = .37).

In a follow up study, Espin and Foegen (1996) examined the relation between reading aloud, maze selection, and vocabulary matching and three criterion measures. Criterion measures were researcher developed and designed to assess comprehension, acquisition, and retention of content information. Results supported the validity of all three measures as indicators of performance on criterion measures ranging from r = .52 to .65. In regression analysis, when reading rate (i.e., correct words per minute obtained by reading aloud) was entered first

vocabulary matching explained a significant amount of the variance when added to the prediction of all three criterion measures. Further, when maze selection was entered first vocabulary matching again explained a significant amount of variance when added to the prediction of all three criterion measures, although reading aloud did not explain any additional variance. Last, when vocabulary matching was entered first maze selection explained little additional variance and reading aloud none. Thus vocabulary matching has been seen as having great potential as a CBM tool for content areas.

Vocabulary Matching

The study by Espin and Foegen (1996) included investigating vocabulary matching was built upon a previous study (i.e., Espin & Deno, 1995) which found vocabulary knowledge the strongest predictor of student performance in content areas (r = .40 to .44). As a result of the Espin and Deno (1995) and Espin and Foegen (1996) studies, additional research has been conducted examining vocabulary matching as a means of estimating students' content area reading skills.

Much of this research has been carried out by Espin and colleagues. For example, Espin, Busch, Shin, and Kruschwitz (2001) examined vocabulary matching as a measure of performance (i.e., an examination of the static score or Stage 1 CBM research). Participants included 58 Grade 7 students in a social studies class. Criterion measures included a teacher created knowledge (pre- and post-) test from the students' classroom and students' performance on the social studies test of the Iowa Test of Basic Skills. Espin et al. (2001) also examined potential differences on vocabulary matching when read by an individual administering the tool or when read by students independently. The stated rationale for examining this difference was a hypothesis that a student's ability to read may impede measuring their level of content

knowledge. Results found no differences for alternate form reliability for administrator- and student-read probes. Administrator-read probes were observed to have alternate form reliability ranging from *r* = .58 to .87 and student-read probes from *r* = .63 to .81. Mean reliability for both types of probes was found to be *r* = .70. Correlations between the vocabulary matching probes and the pre- and posttest ranged from *r* = .59 to .84. Correlations between the vocabulary matching probes and the social studies test of the ITBS ranged from *r* = .56 to .76. Differences between administrator- and student-read vocabulary matching probes was largely not evident.

In a continuation of the Espin et al. (2001) study, Espin, Shin, and Busch (2005) examined vocabulary matching as an indication of progress (i.e., examination of the slope or Stage 2 CBM research). The same criterion measures were again used and differences in administrator and student read vocabulary matching probes was also examined. Using hierarchical linear modeling the mean growth rate estimated for the student read vocabulary matching probes showed an increase of .65 correct matches per week. The mean growth rate estimated for the administrator read probes was .22 correct matches per week. Additional analysis revealed only the student read vocabulary matching probes were sensitive enough to reveal interindividual differences in student growth rates.

In addition, more recent research has been done with vocabulary matching by Mooney and his colleagues. For example, Mooney, McCarter, Schraven, and Haydel (2010) examined the technical adequacy of vocabulary matching as a general outcome measure of social studies content knowledge. Group administered vocabulary matching probes showed strong correlations (*r* = .70) overall and moderately strong to strong correlations (*r* = .63 to .76*)* for demographic variables with the state-wide assessment instrument. Differences reported from the state-wide accountability system in terms of gender, socioeconomic status, exceptionality, and ethnicity,

were reflected in the results of the vocabulary matching achievement patterns.

Mooney, McCarter, Schraven, and Callicoatte (2013) described two studies to evaluate the technical features vocabulary matching. Participants were Grade 6 students (N = 377) in the social studies content area. In study one (N = 153) direct correlation between the curriculum-based measures and the state examination were compared. Vocabulary matching (r = .68) was statistically stronger than the other included criterion measures. Additional criterion measures included CBM Maze (r = .38) and Written Expression (r = .20) and the vocabulary (r = .64) and comprehension (r = .62) subtests of the Gates-MacGinitie Reading Tests, Fourth Edition (MacGinitie, MacGinitie, Maria, & Dreyer, 2000). In study two, linear mixed modeling was used to compare the growth rates for Grade 6 students (N = 224). Across the twenty-five vocabulary matching probes, growth rates varied considerably depending on probe and student socioeconomic status (SES). For example, growth rates from the fall to the spring semesters were significantly different (p = 0.001) for all students. In addition, students who did not receive free or reduced lunch (FRL) were observed to have growth rates significantly different from zero and positive on vocabulary matching probes in the fall and spring (0.23 and 0.11 correct matches respectively). Students who received FRL were observed to have estimated growth rates of 0.10 correct matches in the fall and 0.02 correct matches in the spring. Further, the estimated growth made by students receiving FRL in the spring was not significantly different than zero. Despite the findings related to SES, the overall findings from this research supports the use of timed probes of content vocabulary to serve as an indicator of performance in content areas.

In an extension of the research with vocabulary matching to online tools, Mooney, McCarter, Russo, and Blackwood (2013) evaluated the technical adequacy of vocabulary matching. Science students from Grade 5 (N = 106) were administered 20 parallel forms over a

2-week period. Moderately strong to strong correlations between student performance on the online vocabulary matching probes and the science subtest of the statewide accountability test ranged from *r* = .36 to .55. In regards to alternate-form reliability, weak to very strong correlations were found (*r* = .21 to .73).

In addition, recent research examining the technical adequacy of vocabulary matching probes in science continues to support the use of vocabulary matching as a predictor of performance and progress in content area knowledge. Espin, Busch, Lembke, Hampton, Seo, and Zukowski (2013) investigated the use of 10 vocabulary matching probes over 14 weeks with Grade 7 students (*N* = 198). Criterion measures were knowledge pre- and post-tests and the state-wide assessment measure in the science content area. Alternate-form reliability coefficients were strong (*r* =. 64 to .84), and correlations between the vocabulary-matching and criterion measures ranged from *r* = .55 to .76. Results show the technical adequacy of vocabulary matching as an indicator of progress and performance in science.

Maze Selection

Despite the research support for vocabulary matching some have argued that the novel task – requiring fairly substantial amount of teacher effort to create – may be less appealing than the more familiar maze selection task (see Johnson, Semmelroth, Allison, & Fritsch, 2013). Maze selection has also been examined for predicting student performance and monitoring student progress in content areas (Ketterlin-Geller et al, 2006; Twyman & Tindal, 2007). Often referred to as “concept maze” it is a similar task with the text used being directly taken from the content area of interest (i.e., social studies or science textbook) and words that represent key attributes of core concepts from the content area are targeted. Ketterlin-Geller et al. (2006) administered six concept maze probes over a four-week period after a teacher taught a lesson on

related material but no conclusive data were obtained due to the pilot nature of the study. Two of the coauthors of the pilot study followed up examining potential differences between the use of traditional maze selection, concept maze, and a modified concept maze selection where the attributes themselves were used as response options (Twyman & Tindal, 2008). Regarding the attribute maze probes, test-retest correlations were moderately related (r = .38, .48, and .58). Internal consistency for the two attribute maze probes used were low (α = .27 to .52; α = .22 to .31).

Johnson et at. (2013) examined if maze selection, developed from science content, would demonstrate adequate reliability and validity to function as a tool for reading and science benchmarks. In a study of 367 students in Grade 7, participants completed eight maze selection passages over three testing periods. Approximately half of the students in the study also completed a statewide test of science as well. Results found the science content developed maze selection tools to have alternate form reliability from .56 to .80 and concurrent and predictive correlation coefficients of .63 to .67. Johnson et al. (2013) suggested that while many secondary students struggle in content areas they may not receive assistance as benchmarks for performance are not available to suggest such assistance should be provided. Based on the results of their study, Johnson et al. (2013) suggested the use of maze selection for this purpose continue to be researched. While their results support such a suggestion, research regarding students' reading comprehension (i.e., SVT) may also provide insight into how CBM principles can be applied in content areas.

Sentence Verification Technique

The sentence verification technique (SVT) is a test designed to measure reading comprehension. It is based on the theory that comprehension occurs when a reader constructs

meaning from a linguistic string within a text based on prior knowledge (Royer & Cunningham, 1981). Most important for use as a progress monitoring tool, the theory goes on to include comprehension as occurring when a reader can distinguish whether or not an idea is present in a given text regardless of the words used. Thus, a SVT for content areas would include three to four short 12-sentence passages along with an accompanying set of test sentences. Two of the test sentences have the same meaning as the passage and two do not. For those sentences that have the same meaning one is directly taken from the passage and the other is modified only slightly while maintaining meaning. For the two sentences that do not have the same meaning, one is a complete distractor and the other is only slightly modified though it changes meaning. Students are to read the passage and then indicate whether the accompanying sentences maintain the same meaning as the passage by indicating "yes" or "no." It may be true that a SVT is a more appropriate metric for monitoring progress in content areas as vocabulary knowledge is necessary for being able to successfully complete a SVT.

A primary reason why SVT may be an appropriate measure for monitoring progress in content areas is unique contribution to understanding students' overall reading abilities. Marcotte and Hintze (2009) found that SVT contributed significantly to an overall model of reading comprehension after controlling for oral reading fluency. Participants included 111 students in Grade 4 which included 55 females and was 56% Hispanic; 37% White; 4% African American; and 3% Asian, American Indian, or undisclosed. In addition, 80% of the participants received free and reduced lunch and 30% had a first language other than English.

Marcotte and Hintze (2009) examined four tools commonly used for formative assessment practices in their study as predictor variables. Two criterion measures were also used. The predictor variables included: Three grade level oral reading fluency passages of

approximately 350 words, concurrently collected retell fluency with each oral reading fluency passage, a SVT assessment that included four passages with 16 test sentences each ranging from third to fifth grade in difficulty level, a written retell fluency measure where students silently read a grade level passage for five minutes before writing everything they could remember for an additional five minutes, and a standard fourth grade maze selection passage. Criterion measures included The Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001) and the English-Language Arts section of the Massachusetts Comprehensive Assessment System (MCAS). Marcotte and Hintze (2009) hypothesized that the tools of reading comprehension (i.e., the predictor variables in their study) would account for additional variance beyond oral reading fluency.

Results were that the relation between overall reading proficiency (i.e., the GRADE) and the tools was moderately strong (r = .56 to .67). SVT, along with oral reading fluency and maze selection accounted for approximately 40% of the variance in performance on the GRADE and written retell fluency accounted for 31%. SVT had a medium relation with the other predictor variables, the strongest being with maze selection and the lowest with retell fluency.

Using a multiple-regression analysis, Marcotte and Hintze (2009) found significant results suggesting that approximately 57% of the variance in overall reading ability can be accounted for by oral reading fluency and reading comprehension. Further, maze selection, SVT, and written retell were found to contribute significantly to the overall model of reading after oral reading rate was controlled. In regard to the MCAS, Marcotte and Hintze (2009) found that using a multiple-regression model with oral reading fluency, maze selection, SVT, and written retell as predictors accounted for 66% of the variance in student performance.

Further, SVT may be a better proxy for a students' reading than reading aloud. For

example, Denton et al. (2011) found that the area under the curve for a group administered SVT, as measured by the TOSREC (Wagner & Torgensen, 2010), had the best classification accuracy regarding predicting a desired level of achievement on a state accountability test compared to multiple silent reading fluency measures. Indeed, Denton et al. (2011) found the accuracy of the SVT to be on par with that of tools used to measure oral reading rate.

Limitations of Current CBM Metrics in Content Areas

Despite the potential of these metrics, in particular vocabulary matching and the SVT, there are potential issues present with each as well. As already stated, reading aloud from content area text did not explain any additional variance when vocabulary matching was added first to a regression model for explaining comprehension (Espin & Foegen, 1996). Likewise, maze selection did not explain much additional variance when vocabulary matching was entered first (Espin & Foegen, 1996).

Further, regarding maze selection, Kendeou, Papadopoulos, and Spanoudis (2012) asserted the task functions more as a silent reading fluency measure than one of reading comprehension as students can accurately answer items without considering prior knowledge in the passage. In addition, the common metric of correct restorations used in maze selection has been questioned. Parker et al. (1992) noted that given a student is provided with three choices the likelihood of their correctly responding is nearly 33%. When the nature of distractors is considered (i.e., that one is grammatically incorrect and another is unrelated to the passage's content) Parker et al. (1992) suggested the likelihood of a student correctly responding is close to 66% based on chance alone.

Regarding vocabulary matching, it can be challenging to develop appropriate distractors as well as to develop the task itself as noted by Johnson et al. (2013). In addition, vocabulary

matching may be a measure of one's reading ability with a limited loading on one's knowledge of content being present. As a proxy for a student's knowledge regarding science content this is troublesome. Further, the score obtained by students on vocabulary matching (i.e., number of correct matches) does not allow one to calculate a fluency score. A fluency score is not able to be calculated due to ceiling effects whereas as a high number of students completing the task prior to the allotted time ending results in not being able to determine rate of completion. Without an indication of rate of completion it is not possible to use students' performance to observe response to instruction as it fails to be to change. Vocabulary matching may be an appropriate tool for assisting with screening decisions, or as an indication of subskill mastery. However, by not being sensitive to change, results from vocabulary matching are unable to be used as formative assessment for making progress decisions and guiding instructional changes.

Further, as it has primarily been used as a measure of reading comprehension, prior knowledge is a confound when using SVT to assess students' content knowledge. That is, it is not possible to determine if a student responded correctly to questions related to SVT because of his/her reading of the passage or because the knowledge he/she brought regarding the topic prior to reading. Further, as students are responding to something they have just read results from SVT are too proximal and likely not a good indicator of overall student skill. Also, proximity is a challenge regarding the passage used for SVT as well. That is, the passage can only cover a small component of the curriculum students are taught. This makes it possible that students' performance on one particular passage would not reflect their knowledge on topics related to others passages, or the content of interest overall. In addition, the creation of a SVT passage is a process requiring detailed questions in order to ascertain student's understanding of the passage. As such, the SVT is not a tool which can be briefly administered so that less time can be spent on

assessment and more on instruction. Thus, SVT – like vocabulary matching – is missing characteristics lending it to be a useful tool for helping teachers make screening and progress decisions.

Rationale for a New Approach

Given the nature of learning in content areas (i.e., students reading to learn), measuring students' ability to read and respond to content-specific questions, based on academic vocabulary, is likely to provide an estimate of student content knowledge. In particular, observing the rate of students' responses is likely to be useful. This is related to the theory of automatic information processing in reading (LaBerge & Samuels, 1974). According to this theory, a fluent reader decodes text automatically and thus is able to focus cognitive attention on comprehension. On the other hand, beginning and struggling readers do not automatically decode text due to their need to focus attention on decoding and blending. As a result of beginning and struggling readers' attention being focused on these prerequisite reading skills, they have less cognitive attention available to focus on the meaning of text, making comprehension difficult.

Development of knowledge in content areas, including science, can be viewed in a similar light. This is because mastery of factual knowledge such as definitions and concepts is important for proficiency in application or reasoning (Anderson et al., 2001). By using a tool to measure student comprehension of the accuracy of factual sentences, it is possible to predict general science understanding and application skills. The results of such a tool can assist teachers and other educators with effective instructional decision making regarding the efficacy of provided instruction, students' learning of the material, and current resource allocation.

Study Framework

In general, one of two approaches has been taken with research focusing on CBM (Fuchs, 2004). One approach has been identifying robust indicators of student performance. Fuchs (2004) describes this approach as the process of "identifying a task that correlates robustly (and better than potentially competing tasks) with the various component skills constituting the academic domain" (pg. 2). This approach is related to the concept of dynamic indicators of basic skills (DIBS; Shinn, 1989). DIBS is concerned with the measurement of basic skills that are sensitive to growth and can be used to reliability predict student performance on a future meaningful outcome. Oral Passage Reading (OPR) is an example of this approach as it is a good indicator of students' overall reading ability (Wayman et al., 2007). This is due to one reading at an appropriate rate requires an individual to possess adequate skill levels regarding additional subskills (e.g., letter recognition, blending, vocabulary, etc.). OPR uses the WCPM metric which is obtained by taking the number of words a student reads in one minute from connected text and subtracting the number of decoding errors made (Hosp et al., 2007).

A second approach taken in CBM research has been to use sampling from the curriculum over the course of a school year (Fuchs, 2004). When this approach is used, CBM probes are developed to be administered weekly and each probe involves having students respond to items representing all the skills they are expected to learn by the conclusion of the school year. The sampling approach, more commonly used in math, includes equivalent passages of different types of computation items (e.g., addition, subtraction, multiplication, and division of whole numbers) a student would be expected to be able to solve over the course of a school year. The total score, CD, provides an approximation of a student's current overall proficiency regarding the math curriculum for the school year (Fuchs, 2004).

Three stages of CBM research have been identified regardless of the approach taken (Fuchs, 2004). Stage 1 research addresses technical features of the static score; i.e., the adequacy of a CBM metric for measuring students' academic performance at one point in time through correlational studies to show a relation between a CBM metric and a related criterion measure (e.g., a standardized, norm referenced test of academic achievement). Stage 2 research addresses technical features of the slope, i.e., using regression techniques, to establish a relation between an increasing CBM metric and improvement regarding overall competency in a related academic area. Stage 3 CBM research addresses the instructional utility of a CBM metric, i.e., investigating if the use of a CBM metric improves instructional decision-making and student outcomes.

Most CBM research has focused on Stage 1 (Fuchs, 2004). As such, the identification of CBM metrics useful for making screening decisions in reading (OPR), mathematics (CD), and written language (correct writing sequence; CWS) has developed. While not as extensive as Stage 1, Stage 2 research, especially in regards to reading, has shown these metrics to be adequate for assisting with making progress decisions as well. Less clear has been identifying an appropriate CBM metric for content areas such as science. Learning in content areas involves students reading and understanding content-area material, acquiring information from teacher instruction, and retaining the information they acquire (Espin & Foegen, 1996). SV-S is being developed to measure students' reading and understanding of content-area material. The development of the alternate forms being examined in this study is the result of previous research.

Previous Study of SV-S

In beginning to investigate SV-S, Hosp and Ford (In prep) tested over 800 science

statements with 1,846 students in the fall of 2013. As previously mentioned, statements were based on the Iowa Core Science Standards and the Next Generation Science Standards. In addition, statements were also vetted through an expert panel of science researchers and educators. Participating students were in either Grade 7 or 8 and attended one of three junior high schools in a Midwestern city. Regarding demographic characteristics, 48.9% of students were female, 75.6% were white, 33.2% received FRL, 11.2% were students with disabilities, and 3.4% were English language learners. In testing potential statements, the Science test of the Iowa Assessments was used as the criterion measure. Students completed both SV-S items and the science test of the Iowa Assessments concurrently in their homeroom with directions read by either the classroom teacher or school principal. Students were given 35 to 40 minutes to respond to as many statements as possible. The goal of the study was for students to respond to as many statements as possible. Thus administration of SV-S items was only timed for the logistical purpose of schools continuing with their typical day with minimal interference. Further, in order to test as many items as possible, items were chain linked across seven alternate item sets with the first 30 items overlapping and an additional 120 unique items. Unique items were presented as 1 through 120 or 120 through 1 in order to ensure enough completed items for analysis. Thus, 14 alternate forms were created. Figure 1 demonstrates how common items were overlapped. Table 1 demonstrates how common and unique items were distributed across alternate forms.

Statement responses were scored (i.e., correct or incorrect) and entered into a database. Student performance on the Iowa Assessment Science test was obtained and merged into the same database. Next, statements were analyzed using 2pl item response theory to select the items with the highest discrimination index. Then, using the difficulty index to maintain equivalence, items were identified for use in further investigation of SV-S.

Common Items	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
1-15	15 items						
16-30		15 items					
31-45			15 items				
46-60				15 items			
61-75					15 items		
76-90						15 items	
91-105	15						items

Figure 1. Chain Linking Plan Across Seven Sets of Items. Each set contained 30 items in common (C) with other sets (15 items with one set and 15 with a different one).

Table 1

Distribution of Common and Unique Items Across Alternate Forms for Testing SV-S Items

1A	1B	2A	2B	…	7B
C1-15	C1-15	C1-15	C1-15		C76-90
C91-105	C91-105	C16-30	C16-30		C91-105
U1-120	U120-1	U121-240	U240-121		U840-721

Note. C = Common Items; U = Unique Items; N = 129 – 136 students per form.

In addition to testing potential items for SV-S, Hosp and Ford conducted a follow-up study of 331 students in Grade 8 from one of the three junior highs which participated in testing times was conducted to determine the optimal time for administration and the number of items to include. Students were given one of the alternate forms used for testing items and instructed to respond to statements for 5 minutes. At 1 minute intervals, students were asked to circle the item they had most recently completed. Table 2 shows the descriptive statistics for attempted items by students participating in this study. At 3 minutes the minimum number of items attempted by students increased from 5 to 10. Further, when given an additional minute to respond to items the minimum number attempted only increased to 12. Also, less than 20% of students were observed to respond to more than 50 items at the 3 minute mark. As such, it was determined 60 items and 3 minutes would be sufficient for the number of items to include on alternate forms of SV-S for screening decisions and the time to administer the tool.

Based on the results of these two studies the current forms for SV-S were developed (a more detailed description of SV-S is provided in the methods section). Using a modified sentence verification task will require a working knowledge of vocabulary, but in order to

Table 2

Descriptive Statistics for Determining Optimal Number of Items and Administration Time for SV-S

Statistic	1 minute	2 minutes	3 minutes	4 minutes	5 minutes
Mean	9.20	16.50	23.86	31.87	39.33
SD	3.54	5.45	8.45	11.79	15.35
Skewness	1.36	2.14	2.24	2.64	3.19
Kurtosis	3.32	9.55	9.18	12.29	17.19
Minimum	2	5	10	12	15
Maximum	26	51	72	104	150

Note. SV-S = Statement Verification for Science; SD = standard deviation.

correctly respond to statements will also require an understanding of the concept related to the statement. Thus, this tool can be viewed as a hybrid between a typical SVT and vocabulary matching measures.

Current Study of SV-S

Classical test theory (CTT) will be used to investigate the technical adequacy of SV-S. Within CTT, the focus of test development is on whole-test reliability and consideration of test performance is based on an individual's true score plus error (Crocker & Algina; Lord, & Novick, 1968). Further, several reliability coefficients can be calculated within CTT. Various methods for estimating reliability are discussed below. A discussion regarding evidence of validity follows.

Reliability

In regard to tools used to assess learning, reliability refers to consistency of results. The score an individual obtains on a tool is referred to as an obtained, raw, or observed score. Included in this score is some degree of error which can be either systematic or random. Systematic error consistently increases or decreases a score, often due to inappropriate administration practices. For example, if a tool is administered for too short of a time period all individuals may obtain a score lower than their "true score." Given systematic error is chiefly due to inappropriate administration practices, it can be removed via careful construction and administration of a tool.

While the influence of systematic error can be eliminated, random error erratically influences an obtained score. Examples of random error include differences in performance on a tool at various times of the day, due to illness during administration, or because of guessing. Because random error cannot be completely eliminated, an obtained score on any tool to assess learning will include some measurement error. A primary method for estimating the reliability of a tool is with a reliability coefficient. The reliability coefficient is similar in concept to the validity coefficient discussed above, however in this case it refers to "the correlation between two sets of measurements taken from the same procedure" (Gronlund & Waugh, 2009).

There are four basic methods of estimating reliability: Internal-consistency, test-retest, alternate forms, and test-retest with alternate forms. A reliability coefficient is used for all methods of estimating reliability, however inferences cannot be made across the different methods (e.g., a reliability coefficient concerning test-rest reliability does not allow one to make an inference regarding the internal consistency of a tool).

Internal-Consistency

Internal-consistency methods for estimating reliability require only a single administration of a tool. Multiple methods for estimating internal-consistency of a tool are available. However, all provide an estimate of item consistency. For example, the split-half method involves scoring the odd and even items of a tool independently and correlating the two sets of scores to obtain a reliability coefficient. Cronbach's alpha is another common statistic for measuring internal consistency and is calculated from the pairwise correlations between items. Cronbach's alpha is especially useful for tools which contain items from different areas within a single construct (Dunn, Baguley, & Brunsden, 2013).

Test-Retest

The test-rest method for estimating reliability is concerned with the consistency of a tool's results over a given period of time. Thus, the same form of a tool is given to the same group of individuals at two different times.

Alternate Forms

The alternate, sometimes called equivalent, forms method of estimating reliability is concerned with the consistency of a tool's results given different forms. To estimate alternate form reliability, each form is administered to the same group of individuals at the same time period. Forms of the tool consist of a different sample of items designed to measure the same construct. When a high reliability coefficient is obtained when estimating alternate forms reliability the two forms can be considered to measure the same construct. When a low reliability coefficient is obtained the two forms are considered to be measuring a different construct.

Test-Retest with Alternate Forms

The test-retest with alternate forms method of estimating reliability is a combination of those previously discussed. Thus, two different forms of a tool are administered with a given amount of time in between. Given this method of estimating reliability considers all possible sources of variation it is the most useful for most purposes (Gronlund & Waugh, 2009). That is, when a high reliability coefficient is obtained using the test-retest with alternate forms method for estimating reliability one can infer that an individual would perform similarly on one set of items at one time compared to a different set of (related) items at another time.

Evidence of Validity

Gronlund and Waugh (2009) note four different forms of evidence of validity: Content-related, criterion-related, construct-related, and consequence of using results. Further, when considering to what degree a tool is valid one must be sure to consider the purpose of the tool (i.e., a tool may have evidence of validity for one purpose but not another). In addition, one must also remain cognizant that validity is not dichotomous and that the validity of a tool increases when evidence is available from multiple sources.

Content-Related

Content-related evidence of validity is of the utmost importance in achievement assessment because of interest is how well a tool is aligned with intended learning outcomes. Of specific importance is the adequacy of the sampling of items intended to represent the larger domain of interest (e.g., using a physics test to infer overall physics knowledge).

Criterion-Related

Criterion-related evidence of validity is obtained via two types of studies. One is a predictive study where performance on a tool is used to predict future performance on another.

This latter tool is considered a criterion. An example of a predictive study would be the prediction of first year college GPA given a student's ACT score. The second type of study is a concurrent study. In a concurrent study, performance on one tool is used to estimate current performance on a criterion.

Gronlund and Waugh (2009) note that the benefit of using a predictive study is clear, but that the purpose of a concurrent study may not be as readily apparent. However, they discuss multiple reasons for their use, including: (a) When a new tool is developed, students' performance on the tool will need to be compared to their performance on an established tool of the same domain, (b) There may be an interest in substituting a tool which is quicker, or less expensive to administer, if the tool provides similar results as one which is more time-consuming or expensive, and (c) Of interest may be whether a tool has potential to be used for predictive purposes. Regardless of the type of study used to obtain criterion-related evidence of validity the relation between performance on two tools is expressed using a correlation coefficient.

When a correlation coefficient is used in a study to obtain criterion-related evidence of validity it is called a validity coefficient. When high performance on both tools of interest are observed a positive relation is said to exist. When high performance on one tool is associated with low performance on another a negative relation is said to exist. A perfect positive relation is expressed with a correlation coefficient of 1.00 while a perfect negative relation is expressed with a correlation coefficient of -1.00. A correlation coefficient of 0 represents no relation between two tools.

Construct-Related

Construct-related evidence of validity is concerned with inferences regarding theoretical traits (e.g., reading comprehension, mathematical problem-solving) based on performance on a

tool. Construct-related evidence of validity can be obtained in numerous ways. This includes content- and criterion-related evidence of validity as well as the internal consistency of a tool (see reliability below). In regard to criterion-related evidence of validity, construct-related evidence of validity can be obtained via evidence of convergent (i.e., a tool of interest correlating with other tools representing a similar construct) and divergent validity (i.e., a tool of interest not correlating with tools representing different constructs).

Consequence of Using Results

As previously mentioned, the concept of validity as it applies to tools to assess learning is concerned with the inferences one can make regarding results from such a tool. Thus, it is prudent to consider what consequences may occur from using results of a tool. Gronlund and Waugh (2009) state that when considering the consequence of using results the influence (e.g., motivation, performance, adverse effects) on students should be given great importance. An additional consideration regarding the consequence of using results is whether or not student learning occurs as a result of a tool being used.

Research Questions

Based on the purpose of this study, and its foundation in CTT, the following research questions have been developed:

1. For students in Grades 7 and 8, what is the internal consistency, alternate form, and test-retest reliability of SV-S?
2. For students in Grade 7 and 8, what is the evidence for criterion- and construct-related validity of SV-S?

CHAPTER THREE

METHODS

Chapter Overview

The purpose of this study is to examine the technical adequacy of alternate forms of SV-S for use in assisting educators with making screening decisions (i.e., identifying which students are likely and unlikely to meet expectations on a meaningful outcome related to science). To conduct this examination, two forms will be tested and appropriate indices of validity and reliability will be calculated. The Science test of the Iowa Assessments (Iowas; Hoover et al. 2003) will function as the criterion measure to examine evidence for criterion related (concurrent) validity. Such a test represents a meaningful outcome as a statewide test for the purpose of accountability. The Reading and Mathematics tests of the Iowas will be used to examine evidence for construct validity. The Reading test will examine evidence for convergent while the Mathematics tests will examine evidence for discriminant validity.

Participants and Setting

Participants for this study were students in Grades 7 (N = 799) and 8 (N = 746) from a small city in a Midwestern state who attended one of the three junior high schools in the city's school district. Students from all three schools participated. A random sample of 33% of students from each participating school was identified prior to analysis (discussed below). The remaining Hold Out set and the created Validation set were compared across demographic variables (e.g., gender, ethnicity, free/reduced lunch, English Learner, student with an Individualized Education Program) to ensure sets were of similar composition regarding these characteristics. The Validation set was then separated from the Hold Out Set and analysis of each set was done independent of the other. Demographic information can be found in Table 3.

Table 3

Demographic Data for Participating Schools

Demographic	Grade 7			Grade 8		
	Total (N = 799)	Hold Out (N = 534)	Validation (N = 265)	Total (N = 746)	Hold Out (N = 498)	Validation (N = 248)
Female	47.7%	48.7%	45.5%	50.7%	49.4%	53.2%
African American	21.0%	20.4%	22.3%	19.4%	18.7%	21.0%
Asian	7.4%	6.7%	8.7%	6.7%	6.0%	8.1%
Hawaiian / Pacific Islander	0.4%	0.4%	0.4%	0.0%	0.0%	0.0%
Hispanic	11.3%	10.3%	13.2%	10.9%	10.8%	10.9%
Native American	1.4%	1.5%	1.1%	1.2%	1.0%	1.6%
White	75.0%	76.6%	71.7%	77.7%	78.9%	75.4%
English Language Learner	6.8%	6.4%	7.5%	5.1%	5.0%	5.2%
Individualized Education Program	12.4%	9.7%	10.2%	11.5%	11.4%	11.7%
Free and Reduced Lunch	35.3%	33.7%	38.5%	30.8%	29.5%	33.5%

Instruments

SV-S. Statement Verification for Science has been developed in conjunction with the Iowa Core Science Standards and the Next Generation Science Standards. Statements reflect key concepts included in the Iowa Core and Next Generation Science Standards. Statements were reviewed by one of three science content experts prior to testing items. Two of these individuals were clinical faculty for a university's science education program, and the third was a science consultant for an area education agency. Each form includes 60 items (46 true, 14 false) and is administered for three minutes. The alternate forms for assisting educators with making screening decisions, being examined in this study, have average discrimination values of 1.091 and 1.100 and average difficulty values of - 0.369 and - 0.374 for Form A and B respectively. SV-S may be group administered, making it feasible to implement in schools.

Iowa Assessments. During the course of the participating schools' regular assessment schedule students completed the Iowas (formerly the Iowa Tests of Basic Skills, ITBS; Hoover et al., 2003) as its high-stakes, annual assessment of student achievement.
Students in grades 7 and 8 were administered both Form F of the Iowas (Level 13 for grade 7 and Level 14 for grade 8; http://itp.education.uiowa.edu/ia/ScopeAndSequence.aspx).

Eight tests comprise the core battery of the Iowas for students in grades 7 and 8. These tests include: Reading, Written Expression, Math, Vocabulary, Spelling, Capitalization, Punctuation, and Computation. A test for Science is also available, but is not included in the core battery. For the purposes of this study (see Data Analysis below) the Reading, Math, and Science tests were of interest.

The Reading test is administered in two parts, each for 30 minutes. Students in grades 7 and 8 are provided 45 and 46 items, respectively, to complete. Items cover seven subtests:

Literary Text, Informational Text, Vocabulary, Explicit Meaning, Implicit Meaning, Key Ideas, and Author's Craft.

The Math test is also administered in two parts for 30 minutes each. Students in grades 7 and 8 are provided 70 and 75 items, respectively, to complete. Items cover five subtests: Number Sense & Operations, Algebraic Patterns & Connections, Data Analysis, Probability & Statistics, Geometry, and Measurement.

The Science test is administered at one time for 35 minutes. Students in grades 7 and 8 are provided 41 and 43 items, respectively to complete. Items cover three subtests: Life Science, Earth & Space Science, and Physical Science.

Procedures

Data collection. In order to conduct the appropriate analysis for examining the technical adequacy of alternate forms of SV-S, students were administered both forms at two separate times. The first administration time occurred the week prior to students taking the Iowa Assessments. The second administration time occurred two weeks after the first, with a one week grace period to allow for logistical concerns of the participating schools. Both administrations of SV-S occurred within a two week window of students taking the Iowa Assessments.

Schools were administered SV-S during their homeroom periods. Scripted directions were provided to ensure standardization and were read by either the classroom teacher or by the principal of the school using video broadcast (with the classroom teacher present with students). Students completed both forms of SV-S at each administration period. Each form included the student's name and state identification number in order to link it to the student's Iowas results. The order of the forms was counter-balanced within and across classrooms.

Data analysis. Completed SV-S forms were scanned and analyzed using the Volkmann

Method. The Volkmann Method uses computer software to scan images of student's responses and compares them to an answer key. During the testing of items this method was developed and was shown to produce accurate results (i.e., a 0% error rate). SV-S data for statistical analysis (described below) were collected prior to, and after, administration of the Iowas.

Hold-Out validation. Hold-out validation (HOV), also referred to as test sample estimation, involves partitioning an initial data set into two sets (Kohavi, 1995). HOV may be considered a type of cross validation, a model validation technique for determining how well results from a statistical analysis generalize to another set of data, although in HOV the data are never crossed over. When partitioning the data, the set used for validation is randomly selected and typically comprises one-third of the original database (Kohavi, 1995). Using this procedure allows for internal replication of results given the results from analysis for both sets (described below) is congruent. Such an observation strengthens the external validity of a study as it, in essence, provides a mechanism for replication of results.

The results of conducting a series of independent samples t-tests for testing the equality of means for the Hold Out and Validation sets for students in Grades 7 and 8 can be found in Tables 4 and 5 respectively. Means were compared for both forms of SV-S across administration times as well as the Reading, Mathematics, and Science tests of the Iowa Assessments.

Reliability. This study also examined reliability of SV-S using the test-retest with alternate forms method. As such, students were administered SV-S Form A and SV-S Form B the week prior to administration of the Iowas. Students were administered both SV-S forms a second time approximately two weeks after Administration One. Thus, test-retest reliability was examined by calculating bivariate correlations between SV-S Form A from Administration One and Two as well as SV-S Form B from Administration One and Two. Alternate form

Table 4

Independent Samples t-test for Testing Equality of Means for Hold Out and Validation Sets in Grade 7

	Hold Out Set (N = 534)		Validation Set (N = 265)				95% CI of the Difference			
Measure	Mean	SD	Mean	SD	Difference	SE	Lower	Upper	*t*	Sig.
A1	17.55	8.48	16.62	8.91	.93	.65	-.35	2.20	1.429	.153
B1	17.26	8.32	17.62	8.51	-.37	.63	-1.60	.871	-0.581	.561
A2	20.79	9.49	19.86	9.13	.92	.70	-.46	2.31	1.313	.190
B2	20.71	9.73	20.64	9.03	.07	.71	-1.33	1.47	0.096	.924
Reading NSS	252.72	45.82	245.49	48.51	7.23	3.51	0.33	14.12	2.058	.040
Mathematics NSS	252.26	35.49	247.91	33.83	4.35	2.63	-0.81	9.50	1.655	.098
Science NSS	248.63	37.63	245.04	35.63	3.59	2.79	-1.86	9.05	1.292	.197

Note. Degrees of freedom = 797; SD = standard deviation; Difference = mean difference; SE = standard error mean; CI = confidence interval; A1 = Form A from Administration One; B1 = Form B from Administration One; A2 = Form A from Administration Two; B2 = Form B from Administration Two; NSS = National Standard Score.

Table 5

Independent Samples t-test for Testing Equality of Means for Hold Out and Validation Sets in Grade 8

	Hold Out Set (N = 498)		Validation Set (N = 248)				95% CI of the Difference			
Measure	Mean	SD	Mean	SD	Difference	SE	Lower	Upper	*t*	Sig.
A1	19.90	9.63	20.96	9.54	-1.06	.75	-2.53	.40	-1.424	.155
B1	20.50	10.17	20.98	9.87	-.48	.78	-2.01	1.06	-.608	.543
A2	23.09	10.54	23.47	10.73	-.37	.82	-1.99	1.25	-.453	.651
B2	22.92	10.28	22.75	9.86	.17	.79	-1.37	1.72	.220	.826
Reading NSS	262.77	57.79	264.25	55.69	-1.47	4.44	-10.18	7.24	-.332	.740
Mathematics NSS	260.90	41.97	262.91	43.27	-2.01	3.30	-8.48	4.46	-.610	.542
Science NSS	264.35	42.24	263.79	44.95	.56	3.35	-6.03	7.14	.165	.869

Note. Degrees of freedom = 744; SD = standard deviation; Difference = mean difference; SE = standard error mean; CI = confidence interval; A1 = Form A from Administration One; B1 = Form B from Administration One; A2 = Form A from Administration Two; B2 = Form B from Administration Two; NSS = National Standard Score.

reliability was examined by calculating bivariate correlations between SV-S Form A and SV-S Form B from Administration One and Administration Two.

In addition to the test-retest with alternate forms method for examining reliability of SV-S, internal-consistency was also explored by examining student performance during the first administration of SV-S. Due to the difficulty of examining the internal-consistency of timed tools (i.e., a large enough sample size of students may not complete enough items for the analysis) the following procedures were implemented.

Examination of internal consistency was twofold. One, the number of items completed by a certain percentage of students was identified across various cut points for Administration One. It was first determined that the percentages of 16%, 50% and 84% would be used as a normal distribution would suggest these appropriate. However, due to the number of students attempting all 60 items, 16% was not appropriate. That is, approximately 35.5% to 40.3% of students attempted all items depending on grade, form, and set. Further, between approximately 44.9% and 54.6% of students attempted 32 items and approximately 83.8% to 92.7% of students attempted 18 items. Table 6 shows the percentage of students completing the number of items on SV-S identified as cut scores. Two, split half reliability estimates and Cronbach's alpha were calculated as indices of internal consistency for each identified cut point (i.e., 60, 32, and 18 items) for students in the Hold Out and Validation sets for Grades 7 and 8. Calculating both split half reliability and alpha was done in an attempt to obtain a more accurate estimate of reliability than if only one technique were used. That is, the split half reliability method is likely to provide an overestimate of reliability as it only considers student performance via one combination (i.e., odd and even items). Alpha, on the other hand, is likely to provide an underestimate of reliability as the technique considers student performance via all possible combinations (with some

Table 6

Percentages of Students Completing SV-S Items at Identified Cut Points

Grade	Set	A1			B1		
		18	32	60	18	32	60
7	Hold Out (N = 534)	86.5%	45.9%	36.0%	84.6%	46.4%	36.0%
	Validation (N = 265)	85.7%	47.2%	36.2%	83.8%	44.9%	35.5%
8	Hold Out (N = 498)	89.8%	53.6%	38.8%	89.8%	54.6%	40.0%
	Validation (N = 248)	89.5%	54.4%	40.3%	92.7%	54.4%	38.7%

Note. SV-S = Statement Verification for Science; A1 Form A form Administration Time One; B1 = Form B form Administration Time Two.

combinations likely to be poorly correlated by chance alone). Further, internal consistency should increase given an increase in the number of items attempted.

Unlike standards for evidence of validity, standards for reliability have been more clearly agreed upon. For example, reliability coefficients for test-retest with alternate forms and internal-consistency methods will be compared to the guidelines for reliability noted by Salvia, Ysseldyke, and Bolt (2007). That is, given the intent to use SV-S to assist with making screening decisions, it will be necessary to compare obtained reliability coefficients to a standard of $r = .80$ or greater when examining test-retest and alternate form reliability of SV-S. In addition, guidelines provided by Marston (1989) are prudent to consider as well. Such guidelines are

congruent with research examining alternate form reliability of CBM tools for content areas. For example, Espin et al. (2001) found alternate form reliability ranging from r = .60 to .81 (mean = .70) for 11 adjacent vocabulary matching measures (i.e., correlating 1 with 2, 2, with 3, etc.). In addition, Mooney et al. (2013) found alternate form reliability ranging from r = .21 to .73, with a median reliability coefficient of .56, across 20 vocabulary matching passages. Further, Johnson et al. (2013) found reliability coefficients ranging from r = .63 to .67 for content maze selection passages. Thus, the results of research examining the use of CBM tools in content areas has found mostly moderate to strong relations.

In examining Cronbach's alpha as an indication of internal consistency guidelines are available from Kline (2000) such as follows: $r = \geq .9$ = Excellent (i.e., can use for high stakes testing), r = .7 to < .9 = Good (i.e., can use for low stakes testing), r = .6 to < .7 = Acceptable, r = .5 to < .6 = Poor, and r = < .5 = Unacceptable.

Validity. To examine evidence of criterion-related validity, bivariate correlations between each SV-S form and the Reading, Mathematics, and Science tests of the Iowa Assessments were calculated (e.g., Form A from Administration One and Form B from Administration One were compared with all three Iowa Assessments tests). Following Fisher's r-to-z transformation on each correlation, Meng's z test (Meng, Rosenthal, & Rubin, 1992) was used to compare the correlations between each SV-S form and the Science test of the Iowas. Evidence for construct validity was also examined. Evidence for convergent validity was examined using the Reading test of the Iowas while evidence for discriminant validity was examined using the Mathematics test. Evidence for convergent and discriminant validity was examined by calculating bivariate correlations between each form of SV-S and the Reading or Math tests.

Establishing the appropriateness of a tool is an ongoing and recursive process (Messick, 1989). Therefore, determining when enough data have been accumulated to establish evidence of validity is unclear. However, Marston (1989) provided guidelines for interpreting the strength of validity and reliability coefficients for CBM research. These guidelines are as follows: Strong relations, $r = \geq .70$; moderate relations, $r = .50$ to .69; and weak relations, $r = \leq .50$. Wayman et al. (2007), in their literature synthesis on CBM in Reading, used these guidelines to review the technical adequacy of CBM tools with a focus on word identification, reading aloud, and maze selection. Thus, their use in this study is useful for examining the technical adequacy of SV-S.

A great deal of research has taken place since the time of the Wayman et al. (2007) synthesis regarding CBM tools in content areas (i.e., vocabulary matching, maze selection). Therefore, it is important to consider the technical adequacy of such tools when judging evidence of validity for SV-S. As such, Espin et al. (2001) observed evidence of criterion-related validity for vocabulary matching for social studies with a statewide test which were moderate ($r = .56$ to .64) for seventh grade students. Further, Mooney et al. (2010) found an overall strong correlation of .70 for vocabulary matching for world history and a statewide test of social studies for Grade 6 students. In addition, Mooney et al. (2013) found a moderate validity coefficient of .68 with vocabulary matching for social studies with a statewide social studies test which was significantly more correlated with the statewide test than other CBM tools (maze selection and written expression).

In science, Mooney et al. (2013) examined technical adequacy of an online version of vocabulary matching with Grade 5 students with a statewide science test. Participants were administered 20 parallel forms over a two week period and correlations for evidence of weak to moderate criterion-related validity ranged from $r = .36$ to .55 (pooled estimate of the common

correlation between the statewide test and forms was found to be r = .45). Also in science, Johnson et al. (2013) found evidence of weak to moderate criterion-related validity with a statewide science test of r = .46 to .65 across four maze selection passages for students in seventh grade.

Thus, the results of this research has found evidence of validity of essentially a moderate relation when examining the use of CBM tools in content areas to assist educators with making screening decisions.

CHAPTER FOUR

RESULTS

Chapter Overview

In this chapter the results of examining evidence of technical adequacy of alternate forms for SV-S for assisting educators with making screening decisions is discussed. As stated above, all analyses were conducted on both the Hold Out and Validation sets for students in Grades 7 and 8.

Prior to discussing the evidence examining technical adequacy of SV-S, descriptive statistics for Forms A and B across administration times and the Iowa Assessments are presented. A series of independent samples t-tests were conducted comparing students in the Hold Out and Validation sets for both grades to ensure sets were similar for both tools.

Next, evidence of reliability was examined via several steps. First, internal consistency was examined using Cronbach's alpha and split half procedures. Second, findings from calculating bivariate correlations to examine alternate form and test-retest reliability are highlighted (i.e., test-retest with alternate forms). Third, the results of paired samples t-tests comparing student performance on Forms A and B across administration times are presented. Fourth, following Fisher's *r*-to-*z* transformation of each correlation, Meng's *z* test (Meng, Rosenthal, & Rubin, 1992) was used to examine whether or not differences in prediction are present for alternate forms of SV-S across administration times.

After presenting evidence of reliability, the evidence of criterion- and construct-related validity is presented. This includes highlighting the relation between SV-S forms across administration times and the Reading, Mathematics, and Science tests of the Iowa Assessments. Evidence of criterion-related validity was examined by comparing the relation between student

performance on SV-S forms across administration times and the Science test. Evidence of construct-validity was examined by comparing the relation of student performance on SV-S forms across administration times to the Reading (convergent) and Mathematics (divergent) tests. In addition, potential differences between the Hold Out and Validation sets on each form, across administration times, was examined regarding prediction to the Reading, Mathematics, and Science tests of the Iowa Assessments.

In addition to the planned analyses described above to answer research questions, two exploratory analyses were conducted. One such analyses consisted of a series of independent samples t-tests to determine if observed differences in correct responses between students in Grade 7 and 8 were significant. The second exploratory analysis involved examining the number of items represented within different science domains as included in the Next Generation Science Standards (i.e., life sciences, physical sciences, earth and space sciences, and engineering and technology) for each SV-S form.

Descriptive Statistics

Descriptive statistics for Grades 7 and 8 are presented in Tables 7 and 8 respectively. Students in the Hold Out and Validation sets for both grades obtained more correct responses during Administration Time Two for both Forms A and B. Further, students performed similarly when comparing the Hold Out and Validation sets for each grade. For example, in Grade 7 students in the Hold Out and Validation sets obtained approximately 17 correct responses on Forms A and B during Administration One while students in the Hold Out and Validation sets for Grade 8 obtained approximately 23 correct responses on Forms A and B during Administration Two.

Table 7

Descriptive Statistics for Alternate Forms of SV-S and Iowa Assessments for Grade 7 Hold Out and Validation Sets

Set	Measure	Mean	SD	Skewness	Kurtosis
Hold Out (N = 534)	A1	17.55	8.48	.79	1.71
	B1	17.26	8.32	.78	1.30
	A2	20.79	9.49	.68	.63
	B2	20.71	9.73	.74	.64
	Reading NSS	252.72	45.82	-.64	2.13
	Mathematics NSS	252.26	35.49	< .01	-.94
	Science NSS	248.63	37.63	-.09	-.76
Validation (N = 265)	A1	16.62	8.91	1.17	2.67
	B1	17.62	8.51	.55	.73
	A2	19.86	9.13	.92	1.50
	B2	20.64	9.03	.61	.60
	Reading NSS	245.49	48.51	-1.30	5.70
	Mathematics NSS	247.91	33.83	.10	-.77
	Science NSS	245.04	35.63	.30	-.61

Note. SV-S = Statement Verification for Science; NSS = National Standard Score; SD = standard deviation; A1 = Form A from Administration One; B1 = Form B from Administration One; A2 = Form A from Administration Two; B2 = Form B from Administration Two.

Table 8

Descriptive Statistics for Alternate Forms of SV-S and Iowa Assessments for Grade 8 Hold Out and Validation Sets

Set	Measure	Mean	SD	Skewness	Kurtosis
Hold Out (N = 498)	A1	19.90	9.63	.54	.64
	B1	20.50	10.17	.67	.92
	A2	23.09	10.54	.65	.57
	B2	22.92	10.28	.64	.70
	Reading NSS	262.77	57.79	-1.41	4.85
	Mathematics NSS	260.90	41.97	.08	-1.14
	Science NSS	264.35	42.24	-.25	-.83
Validation (N = 248)	A1	20.96	9.54	.93	1.57
	B1	20.98	9.87	.78	1.07
	A2	23.47	10.73	.60	.58
	B2	22.75	9.86	.68	.94
	Reading NSS	264.25	55.70	-1.30	4.39
	Mathematics NSS	262.91	43.27	-.78	3.99
	Science NSS	263.79	44.95	-.87	3.51

Note. SV-S = Statement Verification for Science; NSS = National Standard Score; SD = standard deviation; A1 = Form A from Administration One; B1 = Form B from Administration One; A2 = Form A from Administration Two; B2 = Form B from Administration Two.

Each SV-S form, across administration times, was judged for adequate skewness and kurtosis with values between 1.0 to < 2.0 considered questionable and those ≥ 2.0 considered problematic (Tabachnick & Fidell, 2012). In general both SV-S forms demonstrated appropriate skewness and kurtosis. High kurtosis values were observed in some cases, most notably Form A from Administration One for the Validation set in Grade 7.

In general, student performance on the Iowa Assessments, as measured by the National Standard Score (NSS), was similar for the Hold Out and Validation sets for both grades with the exception of the Hold Out set in Grade 7 being observed to have higher NSSs in Reading, Mathematics, and Science compared to the Validation set. In Grade 8, students in the Hold Out set obtained a descriptively higher NSS on Science but a descriptively lower NSS for Reading and Mathematics. As would be expected, when comparing the different cohorts across grade performance was descriptively higher (i.e., higher NSSs were obtained) for students in Grade 8 compared to Grade 7 on the Iowa Assessments. Such an observation was also observed for student performance on SV-S as students in Grade 8 obtained more correct responses on both forms across administration times.

Research Question #1: Reliability

Internal consistency. Examination of internal consistency of performance, using Cronbach's alpha, for students in Grade 7 (Table 9) on both forms from Administration One found excellent, positive correlations for students attempting all 60 items (n = 193, α = .94; n = 199, α = .93 for the Hold Out set and n = 100, α = .95; n = 96, α = .93 for the Validation set respectively). Excellent, positive correlations were also found for students completing at least 32 items (n = 267, α = .92; n = 272, α = .90 for the Hold Out set and n = 135, α = .92; n = 135, α = .92 for the Validation set). For students completing at least 18 items good, positive correlation

Table 9

Internal Consistency Analysis for SV-S Forms from Administration One by Number of Items Completed According to Identified Cut Points for Grade 7

Set	Type of Internal Consistency	A1			B1		
		18	32	60	18	32	60
		(n = 462)	(n = 245)	(n = 192)	(n = 452)	(n = 248)	(n = 192)
Hold Out (N = 534)	alpha	.79	.92	.94	.78	.90	.93
	Split Half*	.65	.85	.85	.66	.86	.93
		(n = 227)	(n = 125)	(n = 96)	(n = 222)	(n = 119)	(n = 94)
Validation (N = 265)	alpha	.76	.92	.95	.80	.92	.93
	Split Half*	.65	.88	.92	.65	.92	.91

Note. SV-S = Statement Verification for Science; A1 = Form A from Administration One; B1 = Form B from Administration One; alpha = Cronbach's alpha.

*$p < 0.001$.

were found for both the Hold Out (n = 447, α = .79; n = 447, α = .78) and Validation (n = 222, α = .76; n = 230, α = .80) sets.

Examination of internal consistency of performance, using the split half method, for Grade 7 students found strong, positive correlations for students attempting all 60 items (n = 193, α = .85; n = 199, α = .93 for the Hold Out set and n = 100, α = .92; n = 96, α = .91 for the Validation set). Strong, positive correlations were also found for students completing at least 32 items (n = 267, α = .85; n = 272, α = .86 for the Hold Out set and n = 135, α = .88; n = 135, α = .92 for the Validation set). For students completing at least 18 items moderate, positive correlations were found for both the Hold Out (n = 447, α = .65; n = 447, α = .66) and Validation (n = 222, α = .65; n = 230, α = .65) sets. All correlations for split half reliability were found to be statistically significant ($p < .001$).

Examination of internal consistency of performance, using Cronbach's alpha, for students in Grade 8 (Table 10) on both forms from Administration One found excellent, positive correlations for students attempting all 60 items (n = 193, α = .95; n = 199, α = .95 for the Hold Out set and n = 100, α = .94; n = 96, α = .94 for the Validation sets). Excellent, positive correlations were also found for the Hold Out set (n = 267, α = .91; n = 272, α = .91) while good, positive correlations were found for the Validation set (n = 135, α = .88; n = 135, α = .90) for students completing at least 32 items. For students completing at least 18 items good, positive correlations were found for both the Hold Out (n = 447, α = .81; n = 447, α = .79) and Validation (n = 222, α = .76; n = 230, α = .76) sets.

Examination of internal consistency, using the split half method, of performance of Grade 8 students found strong, positive correlations for students completing all 60 items (n = 193, α = .92; n = 199, α = .93 for the Hold Out set and n = 100, α = .91; n = 96, α = .93 for the Validation

Table 10

Internal Consistency Analysis for SV-S Forms from Administration One by Number of Items Completed According to Identified Cut Points for Grade 8

Set	Type of Internal Consistency	A1			B1		
		18	32	60	18	32	60
		(n = 447)	(n = 267)	(n = 193)	(n = 447)	(n = 272)	(n = 199)
Hold Out (N = 498)	alpha	.81	.91	.95	.79	.91	.95
	Split Half*	.68	.85	.92	.66	.85	.93
		(n = 222)	(n = 135)	(n = 100)	(n = 230)	(n = 135)	(n = 96)
Validation (N = 248)	alpha	.76	.88	.94	.76	.90	.94
	Split Half*	.67	.83	.91	.65	.85	.93

Note. SV-S = Statement Verification for Science; A1 = Form A from Administration One; B1 = Form B from Administration One; alpha = Cronbach's alpha.

*$p < 0.001$.

set). Strong, positive correlations were also found for students completing at least 32 items (n = 267, α = .85; n = 272, α = .85 for the Hold Out set and n = 135, α = .83; n = 135, α = .85 for the Validation set). For students completing at least 18 items moderate, positive correlations were found for the Hold Out (n = 447, α = .68; n = 447, α = .66) and Validation (n = 222, α = .67; n = 230, α = .65) sets. All correlations for split half reliability were found to be statistically significant ($p < .001$).

Test-retest with alternate forms. Paired sample statistics for Grades 7 and 8 are shown in Tables 11 and 12 respectively comparing student performance on SV-S across forms and administration times. To determine the relation between SV-S Forms for the purpose of examining alternate form and test-retest reliability Pearson product-moment correlations were calculated. A dependent-samples t-test was used to determine significant differences.

For students in the Hold Out and Validation sets in Grades 7 and 8 there were weak to moderate, positive correlations between Forms A and B from Administration One which were not statistically significant (r = .43 to .58, p = .043 to .979). For Administration Two there were also weak to moderate, positive correlations when comparing Form A and Form B which were not statistically significant (r = .39 to .56, p = .139 to .856).

For students in both the Hold Out and Validation sets in Grades 7 and 8 there were moderate, positive correlations between Form A from Administration One and Form A from Administration Two which were statistically significant (r = .50 to .68, $p < .0001$). There were also moderate, positive correlations between Form B from Administration One and Form B from Administration Two which were statistically significant (r = .54 to .65, p = <.0001 to .002).

Research Question #2: Evidence of Criterion Related Validity

Table 13 presents the results of calculating bivariate correlations to examine evidence of

Table 11

Paired Samples Statistics for SV-S Forms Across Administration Times, for Grade 7

Set	Comparison	Correlation	Mean Difference	SD	SE	95% CI Lower	95% CI Upper	t	df	Sig.
Hold Out (N = 534)	**A1 to A2**	.50	-3.24	9.04	.39	-4.01	-2.47	-8.280	533	<.0001
	A1 to B1	.46	.30	8.71	.38	-.45	1.03	.775	533	.439
	A2 to B2	.41	.08	10.45	.45	-.81	.97	.182	533	.856
	B1 to B2	.54	-3.45	8.71	.38	-4.19	-2.71	-9.155	533	<.0001
Validation (N = 265)	**A1 to A2**	.53	-3.24	8.75	.54	-4.30	-2.18	-6.034	264	<.0001
	A1 to B1	.58	-1.00	8.02	.49	-1.97	-.03	-2.031	264	.043
	A2 to B2	.56	-.77	8.48	.52	-1.80	.25	-1.485	264	.139
	B1 to B2	.65	-3.02	7.39	.45	-3.91	-2.12	-6.640	264	<.0001

Note. SV-S = Statement Verification for Science; SD = standard deviation; SE = standard error mean; CI = confidence interval; df = degrees of freedom; Sig. = statistical significance value; A1 = Form A from Administration One; B1 = Form B from Administration One; A2 = Form A from Administration Two; B2 = Form B from Administration Two; **Bold** = test-retest; *Italics* = alternate form.

Table 12

Paired Samples T-Tests for SV-S Forms Across Administration Times, for Grade 8

Set	Comparison	Correlation	Mean Difference	SD	SE	95% CI Lower	95% CI Upper	t	df	Sig.
	A1 to A2	.53	-3.19	9.79	.44	-4.05	-2.33	-7.282	497	<.0001
Hold Out	*A1 to B1*	.43	-.60	10.57	.47	-1.53	.33	-1.272	497	.204
(N = 498)	*A2 to B2*	.39	.17	11.47	.51	-.84	1.18	.332	497	.740
	B1 to B2	.57	-2.42	9.49	.43	-3.26	-1.58	-5.688	497	<.0001
	A1 to A2	.68	-2.50	9.39	.60	-3.68	-1.33	-4.199	247	<.0001
Validation	*A1 to B1*	.43	-.02	9.85	.63	-1.25	1.22	-.026	247	.979
(N = 248)	*A2 to B2*	.40	.72	11.32	.72	-.70	2.13	.998	247	.319
	B1 to B2	.61	-1.77	8.72	.55	-2.86	-.68	-3.197	247	.002

Note. SV-S = Statement Verification for Science; SD = standard deviation; SE = standard error mean; CI = confidence interval; df = degrees of freedom; Sig. = statistical significance value; A1 = Form A from Administration One; B1 = Form B from Administration One; A2 = Form A from Administration Two; B2 = Form B from Administration Two; **Bold** = test-retest; *Italics* = alternate form.

Table 13

Bivariate Correlations Examining Evidence of Criterion-Related Validity for SV-S and the Iowa Assessments

Grade	Set	Administration Time	Form	Reading	Mathematics	Science
7	Hold Out (N = 534)	One	A	.08	.12	.11
			B	.21	.21	.21
		Two	A	.18	.20	.17
			B	.23	.23	.21
	Validation (N = 265)	One	A	.19	.17	.18
			B	.21	.20	.24
		Two	A	.22	.21	.24
			B	.26	.23	.25
8	Hold Out (N = 498)	One	A	.26	.28	.26
			B	.27	.28	.26
		Two	A	.27	.32	.26
			B	.31	.35	.32
	Validation (N = 248)	One	A	.31	.27	.33
			B	.34	.30	.33
		Two	A	.24	.26	.21
			B	.19	.22	.18

Note. SV-S = Statement Verification for Science.

criterion- and construct-related validity for SV-S for students in the Hold Out and Validation sets across grades and administration times. General patterns regarding correlations for each test are described descriptively first, followed by an examination of statistical differences in predictions.

Science. For students in Grade 7, performance on Form B was descriptively more highly correlated with the Science test of the Iowa Assessments than Form A across Administration Times for students in both the Hold Out and Validation sets. For students in Grade 8 such a pattern was less prevalent. That is, performance for students in Grade 8 was descriptively more highly correlated with the Science test for Form B during Administration Two than Form A from Administration Two for students in the Hold Out set, while Form A from Administration Two was descriptively more highly correlated than Form B from Administration Two for the Validation set. The correlation coefficients obtained by students in Grade 8 in the Hold Out and Validation sets for Forms A and B from Administration One were not descriptively different at r = .26 and .33 respectively.

Reading. For students in the Hold Out and Validation sets, performance for students in Grade 7 and 8 on Form B, in general, was descriptively more highly correlated with the Reading test of the Iowa Assessments than Form A. The one exception of this finding was for students in the Validation set in Grade 8 where performance on Form A from Administration Two (r = .24) was descriptively more highly correlated than Form B from Administration Two (r = .19).

Mathematics. For students in Grade 7, performance on Form B was again descriptively more highly correlated with the Mathematics test than Form A across Administration Times for students in both the Hold Out and Validation sets. For students in the Hold Out set in Grade 8, performance on Form B (r = .35) was descriptively more highly correlated with performance on the Mathematics test than Form A for Administration Two (r = .32) while Form B from

Administration Two was descriptively more highly correlated than Form A from Administration Two for students in the Validation set (r = .26 to .22 respectively).

Differences in prediction. Table 14 shows the results from conducting independent samples paired z-tests comparing the two sets on each SV-S form for both administration times for students in Grade 7 and 8. Difference in prediction for the Hold Out and Validation across SV-S forms and administration times were not observed for students in Grade 7. In Grade 8 the performance of students in the Hold Out set was more highly correlated with the Mathematics and Science tests of the Iowa Assessment (z = 1.815 and 1.916 respectively) for Form B from Administration Two.

For students in Grades 7 and 8 the results from comparing the correlation between SV-S forms and the Iowa Assessments tests using Meng's z test (Meng et al., 1992) following Fisher's *r*-to-*z* transformation on each correlation can be found in Tables 15 and 16 respectively.
For students in Grade 7, when comparing Forms A and B from Administration One, Form B was found to be a better predictor than Form A for Reading, Mathematics, and Science (z = 2.922, 2.031, 2.254 respectively) for students in the Hold Out set. No difference in prediction was found for the Validation set. When comparing Forms A and B from Administration Two, no difference in prediction was found for students in either the Hold Out or Validation sets. When comparing Form A from Administration One and Two, Administration Two was found to be a better predictor for Reading and Mathematics (z = 2.329 and 1.874 respectively). No difference in prediction was found for the Validation set. When comparing Form B from Administration One and Two, no differences were found in prediction for either the Hold Out or Validation sets.

For students in Grade 8, when comparing Forms A and B from Administration One, no

Table 14

Correlations for SV-S Forms Comparing Hold Out and Validation Sets for Differences in Prediction to the Iowa Assessments, Independent Samples Paired Z-test

Grade	SV-S Form	Reading	Mathematics	Science
7	A1	-1.486	-0.677	-0.948
	B1	0	0.138	-0.419
	A2	-0.552	-0.138	-0.968
	B2	-0.423	0	-0.560
8	A1	-0.697	0.138	-0.982
	B1	-0.989	-0.280	-0.982
	A2	0.411	0.839	0.678
	B2	1.641	1.815*	1.916*

Note. SV-S = Statement Verification for Science; A1 = Form A from Administration One; B1 = Form B from Administration One; A2 = Form A from Administration Two; B2 = Form B from Administration Two.

**** $p < .002$. *** $p < .01$. ** $p < .05$. * $p < .1$.

Table 15

Correlations between SV-S Forms and Iowa Assessments Tests, Meng's z, for Grade 7

Set	SV-S Form	Reading	Mathematics	Science
Hold Out (N = 534)	A1 to B1	-2.922***	-2.031**	-2.254**
	A2 to B2	-1.091	-0.656	-0.869
	A1 to A2	-2.329**	-1.874*	-1.340
	B1 to B2	-0.495	-0.495	0
Validation (N = 265)	A1 to B1	-0.362	-0.541	-1.090
	A2 to B2	-0.716	-0.356	-0.179
	A1 to A2	-0.499	-0.663	-0.999
	B1 to B2	-1.000	-0.597	-0.201

Note. SV-S = Statement Verification for Science; A1 = Form A from Administration One; B1 = Form B from Administration One; A2 = Form A from Administration Two; B2 = Form B from Administration Two.

**** $p < .002$. *** $p < .01$. ** $p < .05$. * $p < .1$.

Table 16

Correlations between SV-S Forms and Iowa Assessments Tests, Meng's z, for Grade 8

Set	SV-S Form	Reading	Mathematics	Science
Hold Out (N = 534)	A1 to B1	-0.218	0	0
	A2 to B2	-0.854	-0.654	-1.281
	A1 to A2	-0.240	-0.974	0
	B1 to B2	-1.013	-1.791*	-1.519
Validation (N = 248)	A1 to B1	-0.499	-0.491	0
	A2 to B2	0.737	0.594	0.440
	A1 to A2	1.434	0.204	2.454**
	B1 to B2	2.778***	1.479	2.769***

Note. SV-S = Statement Verification for Science; A1 = Form A from Administration One; B1 = Form B from Administration One; A2 = Form A from Administration Two; B2 = Form B from Administration Two.

**** $p < .002$. *** $p < .01$. ** = $p < .05$. * $p < .1$.

difference in prediction was found for students in either the Hold Out or Validation sets. No difference in prediction was found when comparing Forms A and B from Administration Two for students in the Hold Out and Validation sets as well. When comparing Form A from Administration One and Two, no difference in prediction was found for the Hold Out set when samples paired z-tests were conducted comparing the two sets on each SV-S form for both administration times in Grade 7 and 8. Differences in prediction for the Hold Out and Validation sets across forms and administration times were not observed for students in Grade 7. In Grade 8 the performance of students in the Hold Out set was more highly correlated with the Mathematics and Science tests of the Iowa Assessment ($z = 1.815$ and 1.916 respectively) for Form B from Administration Two.

Administration One was found to be a better predictor than Administration Two for the Validation set ($z = 2.454$). When comparing Form B from Administration One and Two, Administration Two was found to be a better predictor than Administration One for the Hold Out set ($z = 1.791$) while Administration Two was found to be a better predictor than Administration One for Reading and Science ($z = 2.778$ and 2.769 respectively).

Exploratory Analysis

As discussed in the overview of this chapter, two additional exploratory analyses were conducted for this study. The first exploratory analysis included examining the observed differences in the number of correct responses for students in Grade 7 and 8. The second exploratory analysis included examining the number of items for each NGSS domain included on Forms A and B was also determined. This analysis was included as the items used to develop alternate forms of SV-S were determined based on discrimination and difficulty values and thus representation across domains was not considered. If SV-S is to be used for assisting educators

with making screening decisions, inclusion of items representing all domains is of value.

Grade comparison. A series of independent samples t-tests were conducted to compare the number of correct responses obtained on Forms A and B of SV-S across Administration Times obtained by students in Grades 7 and 8 (see Table 17). Students in Grade 8 were found to obtain more correct responses on both Forms A and B across Administration Times for both the Hold Out and Validation sets.

For the Hold Out set there was a significant difference on Form A from Administration One with students in Grade 8 (M = 19.90, SD = 9.63) obtaining more correct responses than students in Grade 7 (M = 17.55, SD = 8.48), t(992.12) = -4.155, p = <.0001. On Form B from Administration One there was a significant difference with students in Grade 8 (M = 20.50, SD = 17.26) obtaining more correct responses than students in Grade 7 (M = 17.26, SD = 8.32), t(961.61) = -5.59, p = <.0001. On Form A from Administration Time Two there was a significant difference with students in Grade 8 (M = 23.09, SD = 10.54) obtaining more correct responses than students in Grade 7 (M = 20.7, SD = 9.49), t(999.48) = -3.68, p = <.0001. On Form B from Administration Two there was a significant difference with students in Grade 8 (M = 22.92, SD = 10.28) obtaining more correct responses than students in Grade 7 (M = 20.71, SD = 9.73), t(1030) = -3.56, p = <.0001.

For the Validation set there was a significant difference on Form A from Administration One with students in Grade 8 (M = 20.96, SD = 9.87) obtaining more correct responses than students in Grade 7 (M = 16.62, SD = 8.91), t(501.92) = -5.317, p = <.0001. On Form B from Administration One there was a significant difference with students in Grade 8 (M = 20.98, SD = 9.87) obtaining more correct responses than students in Grade 7 (M = 17.62, SD = 8.51), t(488.95) = -4.115, p = <.0001. On Form A from Administration Two there was a significant

Table 17

Independent Samples t-test for Testing Equality of Means for students in Grades 7 and 8 Across Hold Out and Validation Sets

		Grade 7		Grade 8				95% CI of the Difference				
Set		Mean	SD	Mean	SD	Difference	SE	Lower	Upper	*t*	*df*	Sig.
Hold Out (Grade 7 N = 534) (Grade 8 N = 498)	A1	17.55	8.48	19.90	9.63	-2.35	.57	-3.46	-1.24	-4.155	992.12	<.0001
	B1	17.26	8.32	20.50	10.17	-3.25	.58	-4.39	-2.11	-5.59	961.61	<.0001
	A2	20.79	9.49	23.09	10.54	-2.31	.63	-3.53	-1.08	-3.68	999.48	<.0001
	B2	20.71	9.73	22.92	10.28	-2.22	.62	-3.44	-1.00	-3.56	1030	<.0001
Validation (Grade 7 N = 265) (Grade 8 N = 248)	A1	16.62	8.91	20.96	9.54	-4.34	.82	-5.95	-2.74	-5.317	501.92	<.0001
	B1	17.62	8.51	20.98	9.87	-3.36	.82	-4.96	-1.75	-4.115	488.95	<.0001
	A2	19.86	9.13	23.47	10.73	-3.60	.88	-5.34	-1.87	-4.084	486.38	<.0001
	B2	20.64	9.03	22.75	9.86	-2.11	.83	-3.75	-0.47	-2.532	511	.012

Note. CI = confidence interval; SD = standard deviation; Difference = mean difference; SE = standard error mean; Sig. = statistical significance value; A1 = Form A from Administration One; B1 = Form B from Administration One; A2 = Form A from Administration Two; B2 = Form B from Administration Two.

difference with students in Grade 8 (M = 23.47, SD = 10.73) obtaining more correct responses than students in Grade 7 (M = 19.86, SD = 9.13), t(486.38) = -4.084, p = <.0001. On Form B from Administration Two there was a significant difference with students in Grade 8 (M = 22.75, SD = 9.86) obtaining more correct responses than students in Grade 7 (M = 20.64, SD = 9.03), t(511) = -2.532, p = .012.

Item domain comparison. A doctoral student in science education was contacted to review Forms A and B of SV-S in order to determine the number of items represented by the NGSS domains on each form. Figure 2 shows that the domains of earth and space science (28 and 24 items on Forms A and B respectively), and physical science (26 and 24 items on Forms A and B respectively), were most represented with the latter being slightly more evenly distributed across forms. Life science was less represented on both forms, although twice as many items representing the domain were identified on Form B (12) than Form A (6). No items representing the engineering and technology domain were identified for either Form A or B.

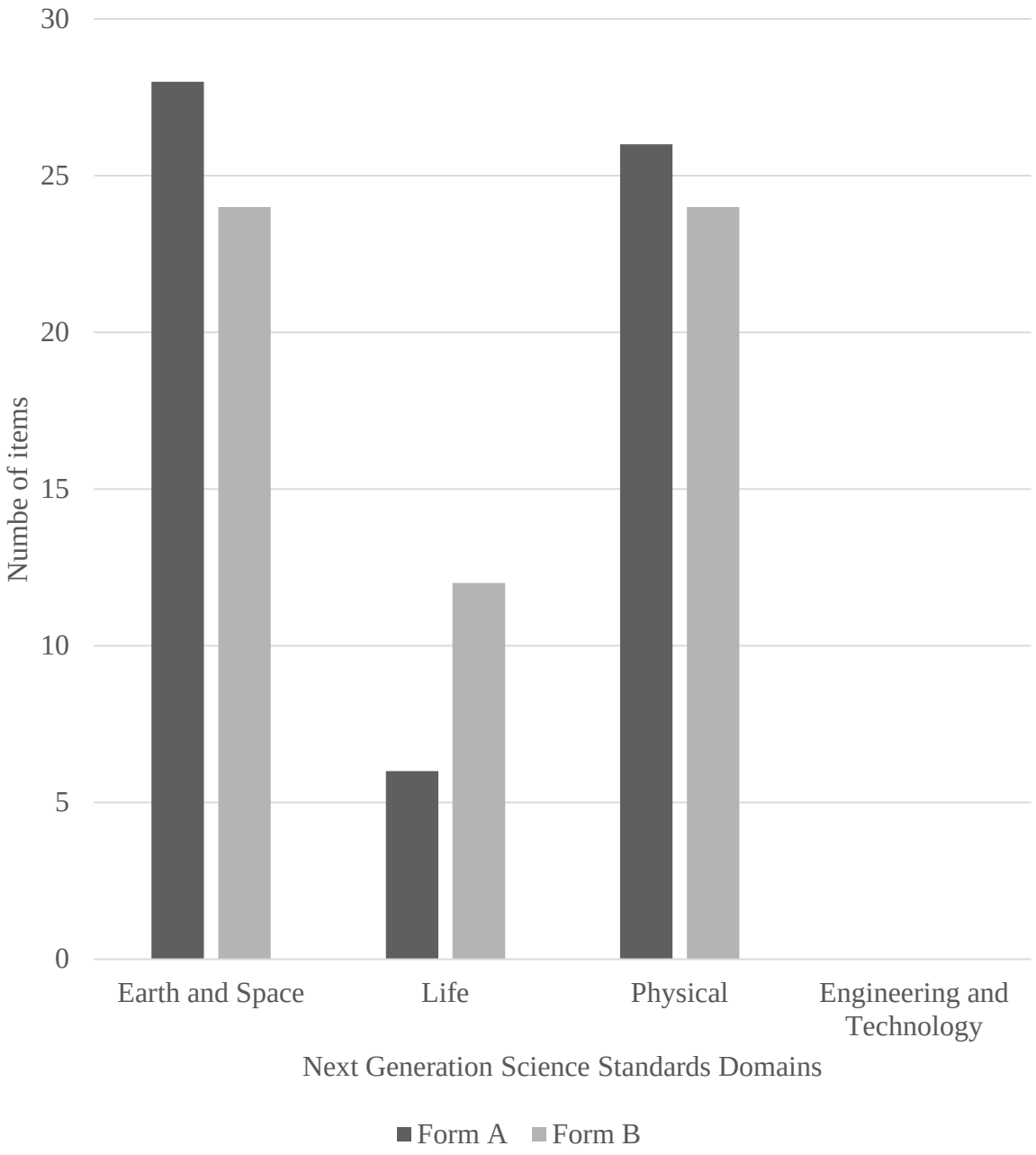

Figure 2. Number of Items on Statement Verification for Science Forms Across Next Generation Science Standards Domains

CHAPTER FIVE

DISCUSSION

Overview

Results from the analyses in CHAPTER FOUR indicate mixed results when examining the technical adequacy of alternate forms of SV-S to assist educators with making screening decisions. Some findings related to reliability met appropriate standards of comparison while those related to evidence of validity were less than desired. Bivariate correlations tended to be lower for both reliability and evidence of validity for SV-S than for those for other CBM tools for use in content areas (e.g., Espin et al., 2001; Espin et al., 2005; Mooney et al., 2010; Mooney et al., 2013; Johnson et al., 2013). However, results were comparable for students in the Hold Out and Validation sets in Grades 7 and 8. Such an observation provides a replication of this study's findings and provides support for narrowing the focus of future research.

The following chapter includes a summary of findings, which presents results from examining reliability and evidence of validity separately; a discussion of the limitations and implications of this study; and areas for future research based on this study's findings.

Summary of Findings

Reliability

This study used several methods to examine the reliability of SV-S. These included examining (a) internal consistency of Forms A and B from Administration One using Cronbach's alpha and the split half method (i.e., even vs. odd items), (b) test-retest reliability of Forms A and B across Administration Times One and Two, and (c) alternate forms reliability for Forms A and B from Administration Times One and Two. The results of each method are summarized below.

Internal consistency. To examine internal consistency of SV-S it was first necessary to

determine appropriate cut scores for items attempted to conduct analyses. That is, due to the timed nature of SV-S, and the need to limit the number of students who finish in order to calculate rate of performance, many students would be expected not to respond to all items. Response to all items is necessary for examining internal consistency. Three cut scores were identified (see Table 6) with one being all 60 items. For Grades 7 and 8, considering both the Hold Out and Validation sets, 35.5% to 40.3% of students responded to all 60 items across forms and administration times. For Grades 7 and 8, given the same considerations, 44.9% to 54.6% and 83.8% to 92.7% of students responded to 32 and 18 items respectively.

For each identified cut score, both Cronbach's alpha and the split-half method of reliability were calculated. As would be expected, an increase in internal consistency was observed given an increase in the number of items attempted by students. Using the standards of comparison set forth by Kline (2000) good, positive correlations were found for 18 items attempted and excellent, positive correlations for 32 and 60 items attempted for students in Grades 7 and 8 on both Forms A and B across sets. This pattern was also observed using the spilt-half method of reliability. Using the standards set forth by Marston (1989), moderate, positive correlations were found for 18 items attempted and strong, positive correlations for 32 and 60 items attempted.

Alternate forms. To examine alternate form reliability, student performance on Forms A and B were correlated for Administration One and Two. For students in Grades 7 and 8 weak to moderate, positive correlations were found for Administration One in both the Hold Out and Validation sets. For Administration Two, weak to moderate, positive correlations were also found for Grades 7 and 8 across sets. Statistical significance was not found when comparing student performance on alternate forms during each administration time.

Test-retest. To examine test-retest reliability, student performance on Forms A and B were correlated across administration times. For students in Grades 7 and 8, moderate, positive correlations were found for Form A from Administration One and Two in both the Hold Out and Validation sets. For Form B, moderate, positive correlations were also found for Grades 7 and 8 across sets. Statistical significance was found when comparing student performance on single forms across administration times.

Evidence of Validity

This study also examined evidence of criterion-related validity via several means. Evidence of construct-related validity was investigated by examining the relation between SV-S and the Science test of the Iowa Assessments. Further, evidence of convergent validity was investigated by examining the relation between SV-S and the Reading test of the Iowa Assessments. Given SV-S is intended to measure, in part, students' ability to read and comprehend science content it was expected the relation between SV-S and the Reading test would be similar compared to its relation with the Science test, but not quite as strong. Evidence of divergent validity was also investigated by examining the relation between SV-S and the Mathematics test of the Iowa Assessments. Given the nature of SV-S as a test of reading content information it was expected the relation between SV-S and the Mathematics test would not be as strong compared to its relation with the Science and Reading tests.

Below, the relations between SV-S and the Science, Reading, and Mathematics tests of the Iowa Assessments are summarized. A summary regarding differences in prediction is also provided for the purpose articulating whether the relation between SV-S and the Science test differs from its relation to the Reading and Mathematics tests.

Construct. Weak, positive correlations were found for both forms of SV-S and the

Science test of the Iowa Assessments for Grades 7 and 8 across administrations times for students in the Hold Out and Validation sets.

Convergent. Weak, positive correlations were also found for both forms of SV-S and the Reading test of the Iowa Assessments for Grades 7 and 8 across administrations times for students in the Hold Out and Validation sets.

Divergent. Weak, positive correlations were also found for both forms of SV-S and the Reading test of the Iowa Assessments for Grades 7 and 8 across administrations times for students in the Hold Out and Validation sets.

Differences in prediction. Given weak, positive correlation were found regarding the relation between SV-S and each test of the Iowa Assessments involved in this study it is not surprising that no pattern regarding differences in prediction were noted. However, in general, no differences were found in prediction for students in the Hold Out and Validation sets for Grades 7 and 8. Such an observation lends support to the relation between SV-S and the Iowa Assessments to be a weak, positive one regardless of the specific test considered.

Exploratory Analyses

Two exploratory analyses were conducted for this study based on observations of results. The first was a comparison of student performance across grades. The second was categorizing SV-S items on Forms A and B in regards to NGSS domains.

Grade to grade comparison. Students in Grade 8 were observed to obtain more correct responses than students in Grade 7 on Forms A and B across administration times in both the Hold Out and Validation sets. The difference between grades was found to be statistically significant. Such an observation provides evidence that more instruction (i.e., having completed another year in school) facilitates higher student performance.

Item domain comparison. Two important findings resulted in examining the number of items related to NGSS domain on each SV-S form. One, was noting a lack of items related to the engineering and technology domain were present on either Form A or B. Two, was the realization that Form A contained half the number of items related to life science as Form B.

Prior to discussing inferences from these results, as well as implications given prior research in this area, potential limitations of this study are presented. This is necessary in order to provide context and caution in interpreting the findings.

Limitations

While results from examining the technical adequacy of SV-S provide promise, several limitations must be noted. One, while the diversity of participants was greater than common for the state in which data were collected, it was not representative of national demographic data. As such, generalizability of results to all students in Grade 7 and 8 should be made with caution.

Further, while evidence of criterion-related validity is important in developing tools for educational decision making, others sources of evidence are needed. Thus, only using the Iowa Assessments to examine evidence of validity is a second limitation of this study. Due to the nature of the development of SV-S (i.e., use of state standards to develop items, expert review of items) a degree of evidence of content- and construct-related validity are present regarding the forms examined in this study. However, additional investigation is needed.

A third limitation of this study is the finding that not all NGSS domains are represented equally across the alternate forms for screening. Of particular concern is the lack of items related to the engineering and technology domain on either form. This suggests that further research regarding item development is needed with SV-S.

A fourth limitation of this study is that analyses were not conducted disaggregating by

student demographic characteristics. That is, given the nature of SV-S as a tool requiring the reading of English it is reasonable to assume students identified as ELs or some students with Individualized Education Programs (i.e., those receiving special education services to remediate reading difficulties) would be observed to obtain fewer correct responses than students who are reading English appropriately at grade level. Research related to providing evidence that students who read well also perform better on SV-S would strengthen the validity of the tool. Further, such research would also allow for insight into likely differences in student performance for these groups compared to the typical performance observed in this study. An additional student demographic comparison not examined in this study which could prove illuminating is potential differences by gender. Girls have been observed to be less likely to remain interested in science and math as they progress through middle school (Orenstein, 1994). This loss of interest has been noted to coincide with a loss of confidence regarding skills related to both subjects, resulting in fewer science and math classes being taken by girls, subsequently contributing to a growing gap in performance between girls and boys (Sanders & Nelson, 2004). Thus, differences in student performance on SV-S by gender are possible. However, despite this limitation, and those discussed above, the findings do have implications and provide direction for future research.

Implications

The National Academy of Sciences (2007) has called for the improvement of science education in United States public schools. This call has been influenced by the prominence of the STEM disciplines in today's society, making it necessary for students to learn the associated knowledge and skills for understanding and using scientific, and related, principles in their everyday life. Given the importance of students acquiring STEM related knowledge and skills it

is prudent for schools to not only provide instruction with a focus on student learning but also measure whether or not students have obtained such knowledge and skills.

Whether or not students have obtained the knowledge and skills determined appropriate is an example of an outcome decision (Hosp et al., 2014). In this case, such a decision is summative in nature (i.e., did the student meet our expectations?). Such a decision is necessary as, at some point, instruction must pause in order to decide if a student has learned what was taught (e.g., did Sam learn all his letter sounds by the end of kindergarten?).

In addition to making outcome decisions, teachers and other educators also make screening decisions (Hosp et al., 2014). Screening decisions, using empirically derived benchmarks, help predict whether students are likely to meet expectations on future outcomes. That is, given how many letters Sam is able to identify at the beginning of kindergarten is it likely he will learn them all (i.e., meet the expectation) by the end of the year? The use of screening decisions can help to identify students who require additional instruction in order to meet future expectations. Used in such a manner, screening decisions are formative in nature.

CBM is a measurement technology system (Hosp et al., 2007) that includes many tools which can be used by educators for making screening decisions. Several CBM tools have been shown to possess the necessary technical adequacy (i.e., reliability and evidence of validity) to be used for making screening decisions (e.g., OPR, maze selection, math computation). However, the research regarding CBM tools for content areas, such as science and social studies, (e.g., Espin et al., 2001; Mooney et al., 2010; Johnson et al., 2013) is still developing.

In an effort to move the research forward in this area, this study examined the technical adequacy of alternate forms of a new CBM tool to assist educators with making screening decisions, Statement Verification for Science (SV-S). Two major theses were present regarding

the development of SV-S. One was the notion that mastery of factual knowledge (e.g., definitions) is important for proficiency in application and reasoning (Anderson et al., 2001). The second was the desire to develop a CBM tool in content areas that allowed for the calculation of rate of performance. This desire was influenced by the theory of automatic information processing in reading (LaBerge & Samuels, 1974). This theory states that rate of performance, or automaticity, is an indication of mastery.

This study found evidence to support SV-S as a reliable (i.e., consistent) tool for assisting educators with making screening decisions about students' science content knowledge. For example, results found both forms of SV-S to have good to excellent internal consistency (α = .88 to .95) for students completing at least 32 items in Grades 7 and 8. Little research has been reported examining internal consistency for CBM tools designed for content areas. However, the results found in this study far exceed those found by Twyman and Tindal (2008) in their examination of internal consistency of attribute CBM maze probes (α = .27 to .52; α = .22 to .31).

In addition, although studies have found some instances of both test-retest and alternate form reliability coefficients higher than .70 and .80 for vocabulary matching (e.g., Espin et al., 2001; Mooney et al., 2010) and, to a lesser degree, maze selection (e.g., Johnson et al., 2013), the moderate test-retest (r = .50 to .68) and weak to moderate alternate form (r = .39 to .58) reliability found for SV-S in this study is congruent with the overall findings of studies examining CBM tools with content areas.

Thus, while falling short of the r = .80 standard for reliability needed for making individual screening decisions identified by Salvia et al. (2007), the results of this study do align with similar research. In addition, the findings related to reliability of SV-S found in this study

were observed to approach the standard identified by Saliva et al. (2007) for group data decisions (i.e., r = .60). In addition, the test-retest reliability found in this study was certainly affected by the observation that students performed better the second time SV-S was administered. Such a difference could be the result of several factors, including (a) increased familiarity with the task and (b) having recently completed the Iowa Assessments prior to taking SV-S the second time (i.e., a priming effect). Thus, test-retest reliability is likely to increase given additional student exposure to SV-S. Further, the alternate form reliability found in this study was likely affected by differences across forms regarding the number of items related to the NGSS domains represented. Alternate form reliability is likely to be increased with more careful consideration of the content represented in items across forms.

However, in light of this study's findings related to reliability, no educational decision – including those related to screening – should be made considering only one source of information. Data from multiple sources should be considered contextually and weighed appropriately. As such, the use of SV-S for assisting educators with making screening decisions appears to merit further consideration based on this study's findings related to reliability.

However, weak, positive evidence of criterion-related validity was found when examining the relation of SV-S with the Iowa Assessments. Studies that have examined the relation between other CBM tools for content areas with high-stakes statewide tests of accountability have found evidence of criterion-related validity in the mostly moderate to strong range (e.g., Espin et al., 2001; Espin et al., 2013; Mooney et al., 2013). One possible explanation for this observation may be related to differences in how SV-S and the Iowa Assessments have been developed. That is, SV-S was developed with deliberate consideration of the Iowa Common Core Science Standards, and reviewed for alignment with the NGSS upon their release,

presumably making it closely aligned to such standards. In contrast, the Iowa Assessments are intended to measure a broader domain of science knowledge which should not necessarily be expected to align with the Iowa Common Core Science Standards. Reason to suspect this may be the case comes from an examination funded by the Iowa Department of Education to investigate the alignment of the Iowa Assessments with the Iowa Common Core Reading and Mathematics Standards (retrieved from the Iowa Department of Education, on March 4th, 2015 from https://www.educateiowa.gov/sites/files/ed/documents/Iowa%20Report%20Math%20and%20Reading%20October%202013.pdf). Results found evidence of less than desired alignment for multiple grades in both subjects, including instances in Grade 7 and 8. For example, only 24% of the items (17 out of 70) on the Iowa Assessments were aligned with the Grade 7 state mathematics standards and 48% of the items (22 out of 46) on the Iowa Assessments were aligned with the Grade 8 state reading standards. It is plausible that such a lack of alignment is also present for Science given the same process and procedures of test development were used (i.e., consideration of state common core standards and expert review).

Further, the weak, positive correlations between student performance on SV-S and the Science (r = .11 to .33), Reading (r = .08 to .34), and Mathematics (r = .12 to .35) tests of the Iowa Assessments, in general, did not result in differences in prediction. That is, student performance on SV-S did not consistently predict their performance on Science better than their performance on Reading or Mathematics. Indeed the only instance in which a better prediction to the Science test was observed was for Form A from Administration Two compared to Form A from Administration One (r = .21; r = .33, respectively; z = 2.454) for Grade 8 students in the Validation set.

Given the results related to the evidence of criterion-related validity for SV-S it is

important to note using it at this time to assist in making screening decisions would be inappropriate. However, it is important to note student performance results suggest the possibility of practice effects. That is, students in both Grades 7 and 8 obtained more correct responses during Administration Two compared to Administration One across SV-S forms (see Tables 11 and 12) which were statistically significant, largely at the $p < .001$ level. This suggests that as students gain familiarity with SV-S, and increase the number of correct responses they obtain, a more accurate representation of their science content knowledge may be obtained. Improvement in more accurately measuring students' science content knowledge would be expected to lead to a stronger relation with criterion measures of science such as the Iowa Assessments Science test.

Further, evidence for the validity of SV-S may be found in the observation that students in Grade 8 obtained more correct responses than students in Grade 7 during exploratory analyses. Given that SV-S items were developed from the band of standards for Grades 6 through 8 in the Iowa Common Core Science Standards, a difference in performance across grades suggests SV-S may, in fact, be measuring real differences in student knowledge as a result of their receiving instruction as they progress through school. Similar observations of student growth are expected on other CBM tools developed using a sampling approach (Fuchs, 2004) such as math computation where an increase in the number of correct digits a student obtains is expected as he/she progresses through school.

Future Research

Given the limitations and implications of this study, multiple directions for future research are present. First, further item development of SV-S is still needed. This includes developing items more closely related to the NGSS domains and ensuring equal representation of

domains across forms. Developing items more closely related to the NGSS domains will facilitate ensuring a better representation of the important scientific, and related, concepts they encompass. In addition, in order to be useful for decision making, alternate forms should contain an equal number of items across domains for the benefit of likely increasing reliability.

Second, future research of SV-S should include examining its relation with other criterion measures. As highlighted above, given the purpose of the Iowa Assessments, and the nature of its development, it is not entirely surprising that a less than desired relation with SV-S was observed. However, in order to be useful for assisting educators with making screening decisions it is paramount that strong, positive relations are observed between SV-S and other, appropriate, measures of student science knowledge (e.g., other state high-stakes accountability tests, student grades, teacher ratings of science knowledge). Establishing such a relation with the Iowa Assessments is also crucial, in particular the Next Generation Iowa Assessments which are being developed to be more closely aligned with the Iowa Common Core.

The relation between SV-S and the Iowa Assessments (and therefore other criterion measures) would likely be strengthened with improved item development. Another means to improve this relation, and a third area for future research, is to examine additional metrics other than correct responses. Given the possibility that students could obtain correct responses on SV-S fairly simply by guessing or randomly indicating “yes” or “no” for a statements, metrics which take into account guessing should be examined. Research examining the use of metrics other than selection of the correct response has been used in maze selection to address this issue (Brown-Chidsey et al., 2003; Deno et al., 2002). Results suggest metrics which take into account incorrect responses (e.g., correct minus incorrect, correct minus ½ incorrect) can improve technical adequacy.

Further, while this study examined the use of correct responses as a static score to begin the investigation of Stage 1 CBM research (Fuchs, 2004) with SV-S, research investigating Stage 2 (i.e., slope) is also necessary. As such Stage 2 research represents a fourth area for future research. Investigating the slope of student performance on SV-S will be necessary for determining whether or not the tool is sensitive enough to measure increase in students' science knowledge.

Conclusions

This study provides evidence that using correct responses on SV-S possesses potential as a tool to assist educators with making screening decisions regarding students' science content knowledge; however work remains to be done. Overall, the reliability of SV-S appears to be a strength of the tool to base further development. Several directions for future research of SV-S should be pursued. Among these directions are the need to focus on improving evidence of validity as well as extending results to additional grades and examining additional metrics. Future research should also extend to examining the slope of student performance on SV-S for technical adequacy as well.

REFERENCES

Alvermann, D. E., (2002). Effective literacy instructional for adolescents. *Journal of Literacy Research, 34,* 189-208.

American educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Raths, J., & Wittrock, M. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Pearson.

Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding Curriculum-based Measurement of Oral reading fluency (CBM-R) decision rules. *Journal of School Psychology*, 51, 1-18.

Bell, B., & Baker, R. (Eds.). (1997). *Developing the science curriculum in Aotearoa New Zealand*. Auckland: Longman.

Beyers, S. J., Lembke, E. S., & Curs, B. (2013). Social Studies Progress Monitoring and Intervention for Middle School Students. *Assessment for Effective Intervention, 38*, 224-235.

Biancarosa, C., & Snow, C. E., (2006). Alliance for Excellent Education. *Washington, D.C.*

Black, P. (1993). Formative and Summative Assessment by Teachers. *Studies in Science Education, 21*, 49-97.

Black, P., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. Granada Learning.

Brown-Chidsey, R., Davis, L., & Maya, C. (2003). Sources of variance in curriculum-based measures of silent reading. *Psychology in the Schools, 40,* 363 – 377.

Chall, J. S. (1999). Models of reading. In D. A. Wagner, R. L. Venezky, & B. Street (Eds.), Literacy: An international handbook (pp. 163–166). New York: Garland Publishing.

Clarke, B., & Shinn, M. R. (2004). A Preliminary Investigation In to the Identification and Development of Early Mathematics Curriculum-Based Measurement. *School Psychology Review*, 33, 234-248.

Committee on Prospering in the Global Economy of the 21st Century. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington DC: National Academy Press.

Curtis, C. K. (1991). Social studies for at-risk and with disabilities. In J. P. Shaver (ed.), *Handbook of research on social studies teaching and learning* (pp. 345-356). New York: Macmillan.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston.

Crooks, T. J. (2002). Educational assessment in New Zealand schools. *Assessment in Education: Principles, Policy & Practice, 9*, 237-253.

Data Recognition Corporation. (2013). *Iowa Assessments Mathematics and Reading Alignment Student Reports Grades 3 -8, 10 and 11.* Maple Grove, MN: Retrieved from the Iowa Department of Education website: https://www.educateiowa.gov/sites/files/ed/documents/Iowa%20Report%20Math%20and%20Reading%20October%202013.pdf.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional children, 52*, 219-232.

Deno, S. L. (1986). Formative evaluation of individual student programs: A new role for school psychologists. *School Psychology Review, 15*, 358-374.

Deno, S. L. (1990). Individual Differences and Individual Difference The Essential Difference of Special Education. *The Journal of Special Education, 24*, 160-173.

Deno, S. L., Anderson, A. R., Calender, S., Lembke, E., Zorka, H., & Casey, A. (2002, February). *Developing a school-wide model for progress monitoring: A case example and empirical analysis.* Symposium at the annual meeting of the National Association of School Psychologists, Chicago, IL.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*, 184-192.

Deno, S. L., & Mirkin, P. K. (1977). Data-Based Program Modification: A Manual. Reston, VA: Council for Exceptional Children.

Denton, C. A., Barth, A. E., Fletcher, J. M., Wexler, J., Vaughn, S., Cirino, P. T., & Francis, D. J. (2011). The relations among oral and silent reading fluency and comprehension in middle school: Implications for identification and instruction of students with reading difficulties. *Scientific Studies of Reading, 15*, 109-135.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36-45.

Deshler, D. D., Palincsar, A. S., Biancarosa, G., & Nair, M. (2007). *Informed Choices for Struggling Adolescent Readers: A Research-Based Guide to Instructional Programs and Practices*. International Reading Association, Newark, DE.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*, 399-412.

Espin, C. A., Busch, T. W., Lembke, E. S., Hampton, D. D., Seo, K., & Zukowski, B. A. (2013). Curriculum-Based Measurement in science learning vocabulary-matching as an indicator of performance and progress. *Assessment for Effective Intervention*, *38*, 203-213.

Espin, C. A., Busch, T. W., Shin, J., & Kruschwitz, R. (2001). Curriculum-Based Measurement in the content areas: Validity of vocabulary matching as an indicator of performance in social studies. *Learning Disabilities Research & Practice*, *16*, 142-151.

Espin, C. A., & Deno, S. L. (1993). Content-Specific and General Reading Disabilities of Secondary-Level Students Identification and Educational Relevance. *The Journal of Special Education, 27*, 321-337.

Espin, C. A., & Deno, S. L. (1995). Curriculum-Based Measures for Secondary Students: Utility and Task Specificity of Text-Based Reading and Vocabulary Measures for Predicting Performance on Content-Area Tasks. *Diagnostique, 20*, 121-42.

Espin, C. A., & Foegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children, 62*, 497-514.

Espin, C., McMaster, K. L., Rose, S., & Wayman, M. M. (Eds.). (2012). *A measure of success: The influence of curriculum-based measurement on education*. Minneapolis, MN: U of Minnesota Press.

Espin, C. A., Shin, J., & Busch, T. W. (2005). Curriculum-Based Measurement in the Content Areas Vocabulary Matching as an Indicator of Progress in Social Studies Learning. *Journal of Learning Disabilities*, *38*, 353-363.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education, 41*, 121–139.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). Effects of frequent curriculum-based measurement on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*, 449-460.

Fuchs, L. S., Fuchs, D., Hamlett, C. L. & Stecker, P. M. (1991) Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations, *American Educational Research Journal, 28*, 617-641.

Fuchs, L. S. (2004). The Past, Present, and Future of Curriculum-Based Measurement Research. *School Psychology Review, 33*, 188-192.

Gonzales, P., Guzmán, J. C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., & Williams, T. (2004). Highlights from the Trends in International Mathematics and Science Study (TIMSS), 2003. NCES 2005-005. *US Department of Education*.

Good III, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, *5*, 257-288.

Gronlund, N. E., & Waugh, C. K. (2009). *Assessment of Student Achievement*. Upper Saddle River, NJ: Pearson.

Hattie, J. (1999). Influences on student learning. *Inaugural lecture given on August*, *2*, 1999.

Hattie, J. A. (2009). Visible learning: A synthesis of 800+ meta-analyses on achievement. *Abingdon: Routledge*.

Hosp, J. L., & Ardoin, S. P. (2008). Assessment for instructional planning. *Assessment for Effective Intervention*, *33*, 69-77.

Hosp, J. L., Hosp, M. K., Howell, K.W., & Allison, R. (2014). *The ABCs of Curriculum-Based Evaluation: A Practical Guide to Effective Decision Making*. New York: Guilford Publications.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABCs of CBM: A Practical Guide to Curriculum-Based Measurement*. New York: Guilford Publications.

Hosp, J. L., & Suchey, N. (2014). Reading assessment: Reading fluency, reading fluently, and comprehension-Commentary on the special topic. *School Psychology Review, 43*, 59-68.

Hosp, J. (2011). Using assessment data to make decisions about teaching and learning. In K. Harris, S. Graham, & T. Urdan (Eds.), *APA Educational Psychology Handbook.* (Vol. 3, pp. 87-110). Washington DC: American Psychological Association.

James, M., & Gipps, C. (1998). Broadening the basis of assessment to prevent the narrowing of learning. *Curriculum Journal*, *9*, 285-297.

Johnson, E. S., Semmelroth, C., Allison, J., & Fritsch, T. (2013). The technical properties of science content maze passages for middle school students. *Assessment for Effective Intervention, 38,* 214-223.

Kamil, M. L. (2003). *Adolescents and literacy: Reading for the 21st century*. Washington, DC: Alliance for Excellent Education.

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22,* 354-367.

Ketterlin-Geller, L. R., McCoy, J. D., Twyman, T., & Tindal, G. (2006). Using a concept maze to assess student understanding of secondary-level content. *Assessment for Effective Intervention, 31,* 39-50.

Kline, P. (2000). The Handbook of Psychological Testing. London: Routlege. In: Clarkcarter, D. (2004) Quantitative Psychological Research. East Sussex: Psychology press.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence, 14,* 1137-1145.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 293-323.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford: Addison-Wesley.

Parker, R., Hasbrouck, J. E., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *The Journal of Special Education, 26,* 195-218.

MacGinitie, W. H. (2000). *Gates-MacGinitie reading tests*. Itasca, IL: Riverside.

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology, 47,* 315-335.

Marston, D. (1989). A curriculum-based approach to assessing academic performance: What is it and why do it. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford.

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33,* 193-203.

McMaster, K., & Espin, C. (2007). Technical features of Curriculum-Based Measurement in writing: A literature review. *The Journal of Special Education, 41,* 68-84.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological bulletin*, *111*, 172.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*, 5-11.

Mooney, P., McCarter, K. S., Schraven, J., & Haydel, B. (2010). The relationship between content area general outcome measurement and statewide testing in sixth-grade world history. *Assessment for Effective Intervention*, *35*, 148-158.

Mooney, P., McCarter, K. S., Russo, R. J., & Blackwood, D. L. (2013). Examining an online content general outcome measure: Technical features of the static score. *Assessment for Effective Intervention*, *38*, 249-260.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.

Nese, J. F., Park, B. J., Alonzo, J., & Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessment: Implications for researchers and teachers. *The Elementary School Journal*, *111*, 608-624.

Orenstein, P. (1994). *Schoolgirls: Young women, self-esteem, and the confidence gap*. New York, NY: Anchor Books.

Parker, R., Tindal, G., & Stein, S. (1992). Estimating trend in progress monitoring data: A comparison of simple line-fitting methods. *School Psychology Review*, *21*, 300-312.

Royer, J. M., & Cunningham, D. J. (1981). On the theory and measurement of reading comprehension. *Contemporary Educational Psychology*, *6*, 187-216.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). *Assessment in special and inclusive education* (10th ed.). Boston: Houghton Mifflin.

Salzman, H., & Lowell, B. L. (2007). Into the eye of the storm: Assessing the evidence on science and engineering education, quality, and workforce demand. *Quality, and Workforce Demand (October 29, 2007)*.

Sanders, J., & Nelson, S.C. (2004). Closing gender gaps in science. *Educational Leadership*, *62*, 74-77.

Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. Guilford Press.

Silberglitt, B., Burns, M. K., Madyun, N. I. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools*, *43*, 527-535.

Stecker, P., Fuchs, L., & Fuchs, D. (2005). Using Curriculum-Based Measurement to Improve Student Achievement: Review of Research. *Psychology in the Schools*, *42*, 795-819.

Stecker, P., & Fuchs, L. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice*, *15*, 128-134.

Tabachnick, B., & Fidell, L. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Torgesen, J. K., & Miller, D. H. (2009). *Assessments to guide adolescent literacy instruction.* Portsmouth, NH: Center on Instruction at RMC Research Corporation.

Tindal, G. (1992). Evaluating instructional programs using curriculum-based measurement. *Preventing School Failure*, *36*, 39– 44.

Twyman, T., & Tindal, G. (2007). Extending curriculum-based measurement into middle/secondary schools: The technical adequacy of the concept maze. *Journal of Applied School Psychology*, *24*, 49-67.

Videen, J., Deno, S. L., & Marston, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression* (Vol. IRLD-RR-84). University of Minnesota, Institute for Research on Learning Disabilities.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of silent reading efficiency and comprehension*. Austin, TX: Pro-Ed.

Wayman, M., Wallace, T., Wiley, H., Tichà, R., & Espin, C. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, *41*, 85–120.

Williams, K. (2001). *GRADE: Group reading assessment and diagnostic evaluation*. New York: Pearson.