Fall 2015

Combining quality and curriculum-based measurement : a suggested assessment protocol in writing

Paula Ganzeveld
University of Iowa

COMBINING QUALITY AND CURRICULUM-BASED MEASUREMENT: A SUGGESTED ASSESSMENT PROTOCOL IN WRITING

by

Paula Ganzeveld

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

December 2015

Thesis Supervisor: Professor John Hosp
Associate Professor Suzanne Woods-Groves

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

Paula Ganzeveld

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 December 2015 graduation.

Thesis Committee: _____
John Hosp, Thesis Supervisor

_____
Suzanne Woods-Groves, Thesis Supervisor

_____
Youjia Hua

_____
Allison Bruhn

_____
Catherine Welch

To Mark, Alexa, and Camden

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. John Hosp for his endless patience, support, and attention to detail. His knowledge of this topic was instrumental in leading me to this line of research, and he supported me through the entire process- even as he was living across the country. Additionally, I would like to thank Dr. Suzanne Woods-Groves were volunteering to step in as a second chair of my committee and keep me grounded. The other members of my committee, Dr. Youjia Hua, Dr. Catherine Welch, and Dr. Allison Bruhn, kept me focused on the big ideas of assessment.

I would also like to thank the administrators, teachers, and children of the Center Point Urbana School District for assisting me in completing this project. I would especially like to thank the second-grade teachers at Center Point Primary School who gave up their valuable teaching time in order for me to collect the necessary data.

Most importantly, I would like to thank my family. Without the love and support of my husband, Mark, I would not have been able to even start this process- let alone finish. I am also grateful to my children, Alexa and Camden, who have grown up so much during this time. I know that they are proud of my accomplishments, but I am even more proud of them and the wonderful young people they have become. Last, I must thank my parents, Paul and the late Carolyn Cartwright. They taught me the value of an education and the necessity to work hard in life. Those lessons that I learned as a child continue to inspire the work that I do as both an educator and a parent.

ABSTRACT

Curriculum-Based Measures in writing (CBM-W) assesses a variety of fluency-based components of writing. While support exists for the use of CBM measures in the area of writing, there is a need to conduct further validation studies to investigate the utility of these measures within elementary and secondary classrooms. Since only countable indices are used in CBM-W, this study explored the possibility of using an assessment that measured writing quality in conjunction with the CBM metric. To accomplish this, three pieces of data were used in this study. The CBM metrics of total words written, words spelled correctly, correct word sequences, percentage of words spelled correctly, and percentage of correct word sequences were scored from a timed writing passage that second grade students completed. Scores from the district writing assessment that classroom teachers rated using an analytic rubric that focused on quality were also analyzed. Last, a validated writing assessment, the TOWL-3, was used as the criterion measure. Using correlation and regression methods, results indicated that correct word sequences was the best predictor performance on the TOWL-3. Even though the teacher writing assessment correlated with the TOWL-3 at the significant level, adding it to the scores from the CBM-W measures did not significantly increase the validity.

PUBLIC ABSTRACT

Writing is one of the most complex processes taught in our public schools. It requires individuals to use knowledge of spelling, grammar, punctuation, organization, and vocabulary, along with fine motor skills. A challenge that teachers have regarding writing is the assessment of students' writing, as it is more subjective than reading or math. Some writing assessments have been thoroughly studied to determine use in schools. One of those assessments is the TOWL-3. The TOWL-3 is a standardized assessment that cannot be used multiple times, so it is necessary to find writing assessments that can be used multiple times throughout a school year to determine if students are making necessary progress. Curriculum Based Measurement (CBM) is an assessment method that use timed tests to assess student's fluency with the content. Even though the CBM reading assessments are commonly used in schools, the CBM measures for writing do not meet the necessary levels. In an attempt to find an assessment that would increase the appropriateness of CBM, this study added scores from a writing assessment used in a school district that analyze quality components of writing (organization, word choice, ideas, etc.). Second grade students completed both CBM and the district assessment, and the scores were compared to the TOWL-3. Through statistical analysis, it was found that the addition of the quality measure did not significantly increase the appropriateness of the CBM measures with students in second grade.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

Writing is a fundamental aspect of competent communication and literacy in modern societies (Behizadeh & Engelhard Jr., 2011). Proficient writing is a critical component for success in school, across elementary, middle, and high school as students are expected to produce more writing through papers and essay tests as they progress through school (Bradley-Johnson & Lesiak, 1989).

However, poor literacy skills play a role in why many students do not complete high school (Graham & Hebert, 2010). Students that struggle with writing find themselves at a serious disadvantage pursuing some form of higher education, securing a job that pays a living wage, or participating in social and civic activities. Technology innovations, along with globalization and changes in the workplace have increased the need for some form of higher education, including technical or career coursework, two-year, or four-year college. Reading and writing are now essential skills in most white- and blue-collar jobs (Graham & Hebert, 2010). High-level literacy skills are almost a common requirement in the service industry. Almost 70% of salaried employees in finance, insurance, real estate, construction, and manufacturing use writing as part of their jobs (Magrath & Ackerman, 2003). Based on reports from workers, over 80% of blue-collar and 90% of white-collar workers indicate that writing skills are important to job success (Magrath & Ackerman, 2003).

In English classrooms, students are expected to write compositions and complete literary analyses. Writing requirements also extend into other content areas. This might include writing lab reports in science, biographies in history, comparing government systems in civics, and explanations of mathematical problem solving in math courses (Alber-Morgan, Hessler, & Konrad, 2007). Not only does writing allow students to

demonstrate their knowledge, it also supports and strengthens cognitive learning strategies. Writing about content learned provides students with more exposure to the content, which also increases time-on-task (Bangert-Drowns, Hurley, & Wilkinson, 2004). As students link new learning with established constructs, they are able to synthesize information, explore relations and implications, and develop conceptual frameworks. This process also encourages metacognitive comprehension monitoring, a major building block for reading-to-learn and writing-to-learn.

Despite the recognized importance, many researchers and policy-makers perceive that the nation is in a crisis in regards to writing instruction (Behizadeh & Engelhard Jr., 2011). The National Commission on Writing (Magrath & Ackerman, 2003) proposed the need for a "writing revolution". This group confirmed the importance of including writing instruction into every class and across all grades, reinforcing the need for states to examine curriculum standards to increase student achievement in written language.

Performance standards are used to describe desirable outcomes for student skills, outlining what students should be able to accomplish in school. The Common Core State Standards initiative and the National Governors Association (2010) have provided set targets, by grade, outlining the writing skills students need to achieve in order to leave high school prepared for college, vocational training, or work-related endeavors (Costa, Hooper, McBee, Anderson, & Yerby, 2012). The new Common Core State Standards recognize writing as a central strand comparable to reading, and may increase the focus writing is given in classrooms across the country (Applebee & Langer, 2011). A recent call has also been made for greater uniformity between academic standards and evaluation procedures. There are persistent discrepancies between state and national test

results (Jeffery, 2009). Gaps between high proficiency levels reported for state mandated assessments and low proficiency levels reported for National Assessment of Educational Proficiency (NAEP) may result from differences in how proficiency is conceptualized between developers from state and national testing programs (Jeffery, 2009).

Even though writing is included in the standards, the responsibility for acquiring the skills rests largely on the students' shoulders. In elementary grades, the focus is on reading, with almost half the academic day devoted to a core reading program (Calfee & Miller, 2007). In middle school classrooms, the amount of time spent on any aspect of writing instruction is very small (Applebee & Langer, 2011). In English classrooms, only 6.3% of time was spent teaching explicit writing strategies, 5.5% on studying writing models, and 4.2% on evaluating writing. This equates to just over three minutes of instruction related to explicit writing strategies in a 50-minute class.

This need to reform writing instruction is substantiated in policy statements related to NAEP and other policy-oriented research on writing (Graham & Perin, 2007). NAEP is currently the only national test used in the United States. The purpose of NAEP is to regularly assess what American students know and are able to do in various subject areas (Conley, 2005). According to NAEP writing assessments, most students do not have proficient writing skills. Despite recent small improvements in written language, 67-76% of students in grades 4, 8, and 12 have either partial or no mastery of core writing skills (Persky, Daane, & Jin, 2003)). Examining just 12th grade students, only one out of four is a proficient writer (Salahu-Din, Persky, & Miller, 2008). Nationally, high school graduates also struggle in the area of writing. Approximately 40% of high school graduates lack the literacy skills employers seek (National Governors Association, 2005)

and over one-third of high school graduates that took the American College Test (ACT) did not meet the benchmark for college readiness (American College Testing, 2014). This indicates that they are not ready for college-level English composition courses. Not only are there concerns related to writing scores, concerns also exist regarding writing instruction in the educational system.

Actual writing that goes on in classrooms across the United States remains dominated by teacher-controlled tasks. Applebee and Langer (2011) found that typical middle school English classrooms include tasks in which the teacher does all of the composing, and students are left to fill in missing information. This might be by copying directly from a teacher's presentation, completing worksheets and chapter summaries, reproducing highly formulaic essays from a highly structured plan, or writing the particular information that the teacher is seeking. A lack of quality assessment tools in the area of writing may be limiting the implementation of quality instruction.

Little research has also been completed on the subject of writing assessment (Huot, 1990; Medina, 2006). Of the 1,502 articles written from 1999-2004 on writing research, only 7.5% addressed writing assessment (Juzwik et al., 2006). There has been significant public discourse regarding writing during these years, but little research has been conducted on how to assess writing. Most educational professionals believe that students must acquire a certain level of writing ability; little agreement exists regarding what that ability level should be or what constitutes a good measure of writing (Cole, Haley, & Muenz, 1997; Hessler, Konrad, & Alber-Morgan, 2009; Medina, 2006).

Most states conduct summative high-stakes assessments once a year in reading, math, and science, but few state testing systems rely on any direct writing assessment to

measure student achievement in writing (Beck & Jeffery, 2007). Instead, multiple-choice item tests dominate writing tests, usually assessing mechanics and convention (Calfee & Miller, 2007). A few states have begun to use writing samples as part of their high-stakes testing programs. Florida administers the Florida Comprehensive Assessment Test to students in elementary and secondary grades. For the writing portion, students are given a writing prompt and 45 minutes to complete a response. Students are instructed to write a narrative, expository, or persuasive response from a prompt. Trained raters evaluate the writing based on focus, organization, support, and conventions. These measures are attempting to assess writing in the form that students typically write, in essay form. There is now added emphasis on this form of writing.

The Common Core State Standards (2010) now recognizes writing as a central strand, comparable to reading in the teaching of English/Language Arts (Applebee & Langer, 2011). Writing is also addressed within the math, social studies, and science standards as a mode for student to express their knowledge in these areas. If all students are to meet rigorous standards, assessment tools are needed to track students' progress toward those standards to quickly and accurately identify students at risk of failing (McMaster & Espin, 2007). If students are not making adequate progress, interventions are needed to close these gaps. The response to intervention (RTI) model may be used to implement and monitor evidence based interventions with students that are having difficulty meeting the core standards. Even though there is an extensive research base in writing, little research discusses how to frame writing within an RTI model (Saddler & Asaro-Saddler, 2013). Fuchs (2004) stated that curriculum based measurement might be the tool used within RTI frameworks, but additional research is needed.

Another important component of RTI is the assumption that early intervention programs yield more benefits than efforts aimed at remediated problems in later grades (Graham, Harris, & Larsen, 2001). Impaired compositional fluency in the primary grades may serve as the origin of writing problems in later grades (Beringer, Mizokawa, & Bragg, 1991). By identifying a writing procedure to quickly identify students with writing concerns in the early grades, interventions can be implemented to remediate these problems while the interventions are still effective.

Purpose of Study

Even though Curriculum-Based Measurement in Writing (CBM-W) has been cited as including reliable and valid measures of written expression (Marston, 1989), much concern still exists using these measures for screening, progress monitoring, and eligibility decisions (Coker & Ritchey, 2010; Espin, Weissenburger, & Benson, 2004; Espin, De La Paz, Scierka, & Roelofs, 2005; Jewell & Malecki, 2005; McMaster, Xiaoqing, & Pétursdóttir, 2009). Researchers have had difficulty identifying CBM-W measures that are both valid and sensitive. Correct word sequences (CWS) has been found to have strong correlation with holistic measures, but is not as sensitive to growth. Total words written (TWW) is more sensitive, but it is weak in validity, especially in recent studies. Also, the measures do not contain the high predictive validity that is seen in reading and math CBM measures. Shriner and Thurlow (2012) suggest that a combination of measures (both quantitative and qualitative) might be needed to identify students' likely performance on measures relative to state or district standards.

Based on Shriner and Thurlow's recommendation, the purpose of this study was to combine CBM metrics with a teacher rating. Timed writing passages were scored using a variety of quantitative metrics to determine which measure(s) correlated most

highly with the Test of Written Language-3 (TOWL-3; Hammill & Larsen, 1996) Spontaneous Writing subtests. Data were analyzed from the district writing assessment that used a teacher rating from an analytic writing rubric. These scores were also correlated with the TOWL-3 to determine concurrent validity. Writing samples from second grade students were used to develop a scoring procedure that can be used to identify students early that have writing difficulties. The following research questions were addressed in this study:

1. Which CBM metrics (TWW, words spelled correctly [WSC], CWS, percentage of words spelled correctly [%WSC] and percentage of correct word sequences [%CWS]), alone or in combination, best predict writing performance as measured by the TOWL-3?
2. What is the relation between the scores on the district writing assessment and the TOWL-3?
3. What impact does adding an analytic rubric score generated from a classroom writing activity have on the CBM metrics' prediction of TOWL-3 performance

CHAPTER TWO: LITERATURE REVIEW

The Common Core State Standards (2010) now recognizes writing as a central strand, comparable to reading in the teaching of English/Language Arts (Applebee & Langer, 2011). Writing is also addressed within the math, social studies, and science standards as a mode for students' to express their knowledge in these areas. An extensive amount of research has been conducted regarding the building blocks associated with students learning how to write and the skills necessary to be considered a proficient writer. In this chapter I will briefly describe these foundational theories of writing, define the characteristics of struggling writers, and delineate the different writing assessments currently used. The framework for Response to Intervention (RTI) in writing will be explored, including the research on Curriculum-Based Measurement (CBM), measures commonly used to determine effectiveness of interventions.

Foundations of Writing

Writing is a complex interaction of both cognitive and physical processes (Bromley, 2011). It involves hand, eye, right, and left sides of the brain to construct meaning and make connections. Writing is an area that has been well researched, but little consensus exists. There is currently no model or theory of writing that is fully accepted (Olinghouse & Graham, 2009). Between 1970 and about 1990, the discourse surrounding writing focused predominantly on prescriptive text features of model prose written by exemplary writers (Nystrand, 2006). Moffett (1968) was one of the first researchers to view writing as a cognitive process. Using Piaget's model, he developed a K-13 pedagogical sequence of writing development based on increasing levels of abstraction. These levels progressed from record, to report, to generalization and analysis, and ended with speculation. This model of English education moved the focus of

curriculum and instruction away from traditional models of cultural heritage and skills (Nystrand, 2006). This was a call for a reconceptualization of writing rooted in basic research about individual learning and process of mind.

Another source of discussion on writing at this time was the Cambridge Cognitive Revolution at MIT and at Harvard University (Nystrand, 2006). With Chomsky's (1957, 1966, 1968) revolutionary research on linguistics, a Cognitive Revolution occurred in writing, reading, and learning in the 1960s. However, the influence on writing did not occur until at least the mid-1970s (Nystrand, 2006). Two influential articles also appeared in mainstream media in the 1970s that claimed a sharp decline in writing skills of college and university students. "Bonehead Engish" (Stone, 1974) attributed unprepared students to the dramatic increases in remedial freshmen composition courses. The Newsweek cover story, "Why Johnny Can't Write" (Sheils, 1975) went one step further, blaming public schools for neglecting "the basics" and cited too many "creative" methods and "permissive" standards in "open" classrooms.

The 1980s saw researchers using this early research to build writing models. Flower and Hayes (1980) theorized one of the first formal models, delineating both the components and organization of writing processes. This original writing model addressed two aspects of written text comprehension. The first aspect, Understand, described the process by which people come to understand written texts. The second aspect, Attend, described the process in which individuals decide what is most important in the written text. The model also had three major components: (a) Task environment, which includes all the outside factors that influence the writing task; (b) Cognitive processes, which

include planning, translating, and revision, and; (c) Writers long-term memory, which included knowledge of topic, audience, and genre.

Bereiter and Scardamalia (1987) developed a similar cognitive model of writing to explain simple knowledge sharing by novice writers as well as knowledge transformations by expert writers. This model used two components of information: content and discourse. It decomposed writing into 4 main processes: (a) mental representation of the task, (b) problem analysis and goal setting, (c) problem translation between discourse and content components, and (d) resultant knowledge-telling. Bereiter and Scardamalia viewed the act of writing as a recursive problem solving process that helps expert writers think more effectively about a topic. Like Flower and Hayes, Bereiter and Scardamalia described self-regulatory strategies as mental subroutines. They also suggested that these strategies contribute to the development of new linguistic rules.

Some researchers challenged these cognitive models of writing. Nystrand (1982) argued that the relations that define written language functioning are based on the systematic relations that exist in the speech community of the writer. Bizzell (1982) also challenged Flower and Hayes, stating that the model lacked the connection to social context by ignoring the dialectical relationship between thought and language. Last, Faigley (1985) argued that the model did not describe writing within a language community where people acquire specialized kinds of writing competence that enable them to participate in specialized groups.

To address some of these concerns, Hayes (1996) updated the model. The new model had two components: the task environment and the individual. The task environment consisted of a social component, including the audience, the social

environment, and other texts that the writer may read while writing. The physical components included the text that the writer has produced so far and a writing medium. The individual component incorporated motivation and affect, cognitive processes, working memory, and long-term memory.

Bradley-Johnson and Lesiak (1989) categorized these writing skills as mechanics, production, conventions, linguistics, and cognition. Mechanics refers to the ability to form letters, words, numbers, and sentences that are readable and legible. Production includes the number of words, sentences, and paragraphs a student is able to produce to convey ideas and feelings. The rules for capitalization, punctuation, and spelling are included within conventions. Linguistics describes the ability to use varied vocabulary and correct syntax. These skills are closely related to students' oral language development. Cognition refers to the organizational aspect of writing, determining if the writing is logical, sequential, and coherent. These five skill areas are interrelated. If a problem exists in one component, it will likely result in difficulty in one or more of the other components.

It is also theorized that story composition for young writers follow a similar pattern. At the first level, children are able to write about a conflict, but have difficulty developing a resolution (Barenbaum, 1987). As children mature, their plot structure develops. They are able to include elaborate conflicts with resolutions that encompass multiple subplots that interrupt the primary plot line. They are also able to develop characters and dialogue beyond simple descriptions. When students do not develop these skills or develop along the lines of their peers, concern arises.

Since writing depends on high levels of personal regulation, Zimmerman, Bonner, and Kovach (1997) developed a social cognitive model of writing composed of forms of self-regulation: environmental, behavioral, and covert or personal. They suggest that these self-regulatory processes interact reciprocally during writing through an interactive feedback loop. This loop is a cyclical process in which writers monitor the effectiveness of their self-regulatory strategies and then react based on the feedback to either continue the strategy if successful or modify or change it when it is not. This was the beginning of a shift to not only describe what students experience during the writing process, but to proceed to build instructional components based on these theories. An explicit writing process continues to be a strong component of more recent writing research.

The most extensively used self-regulated package, developed by Harris and Graham (1996) is known as self-regulated strategy development (SRSD). During this procedure, students are taught to set goals, self-record their progress by graphing the output of targeted elements, use a mnemonic strategy specific to the writing task, use self-instruction, and self-evaluation of their progress. Not as specific as SRSD, Graves' model of writing (1994) focused on engagement and relevance as basic components of the writing process. His model described the recursive steps of planning, drafting, revising, editing, and publishing for a real audience. This process approach to writing has become widely used in school through the writing workshop format. Atwell's well-recognized work (2002) also supports this use of writing processes within the writing workshop format.

The question "What is writing?" began to take on a life of its own in the 1990s. The focus has become more comprehensive, with writing in all its situated contexts both

in and out of school: writing and technology in the work, writing and culture in communities, and investigations of writing in numerous nonacademic settings (Nystrand, 2006). These models of writing provide a foundation to understand the processes students experience as they proceed through the writing process. This is especially the case for students that struggle in writing.

Struggling Writers

Analyzing the processes included in writing leads to the question, what are the differences between skilled and unskilled writers? Skilled writers are described as goal-directed learners that apply various writing and self-regulation strategies. These strategies include planning, revising, organizing, monitoring, and evaluating (Cole et al., 1997). This aligns with both the social cognitive model and the components of the writing process outlined by Graves (1994). When comparing students with learning disabilities and normally achieving students, the students with learning disabilities wrote fewer words and sentences, but wrote more words per sentence. They also produced fewer words with seven letters or more and had a higher percentage of capitalization and spelling errors (Houck & Billingsley, 1989).

Common factors associated with struggling or unskilled writers can be placed in five categories: overall knowledge of writing; planning; generating ideas; transcription; and revising (Graham & Harris, 2002). The overall knowledge of writing includes understanding the attributes and expectation of different writing genres. Struggling writers may omit parts of a story or have difficulty understanding the need for topic sentences. Planning is an area that struggling writers spend little attention on when constructing their writing. This includes development of goals, organization, and addressing the needs of the reader. Since struggling writers produce little writing, their

papers contain a lack of detail or organized ideas, indicating that the students struggle with generating ideas.

The mechanics of writing can also be described as transcribing words to print (Graham & Harris, 2002). Papers written by struggling writers are often full of capitalization, punctuation, and spelling errors. Handwriting is often a frustrating and time-consuming task. These mechanical difficulties impede the process of writing. By switching attention during composing to these mechanical demands, the writer may forget the ideas they are trying to convey and forget or abandon plans that were developed within their working memory. This alignment to the social cognitive model demonstrates that students are not able to plan while writing and have difficulty developing expressions that precisely fit intentions. In order to assist these struggling writers, a system to accurately identify them is needed. Since written language is a complex set of skills that are developed in a predictable sequence, it should be possible to analyze current skills, identify deficits, plan interventions, and determine effectiveness of the interventions (Penner-Williams, Smith, & Gartin, 2009). This process of observing how students respond to interventions is just one component of Response to Intervention (RTI).

Response to Intervention

To fully understand how the RTI model and CBM measures may work together, it is important to fully review RTI. The goal of RTI is to maximize student achievement. The model integrates assessment and intervention in a multi-level prevention system. Schools use the data from assessments to identify students at risk for poor learning outcomes. There are four essential components of RTI. These components include

screening, progress monitoring, multi-level or multi-tier prevention system, and data-based decision making (National Center on Response to Intervention, 2010).

Many states (e.g. Iowa, Florida, Illinois, Kansas) have adopted Multiple Tiered Support Systems (MTSS). This approach uses the multiple tiered system first developed in the RTI model to provide instruction to all students through a tiered support system. The tiers in this approach are designed to match each student's needs with layers of instruction, immediate feedback, progress monitoring, and ongoing assessment (National Center on Response to Intervention, 2010). Tier 1 is the primary level of prevention with differentiated instruction within the scientifically based core curriculum. Students at Tier 2, the secondary level of support, are at moderate risk for academic failure. At this tier, students receive evidence-based interventions in small groups. Students at Tier 3, known as tertiary prevention, are at a high risk for failure.. These interventions at this tier are even more intensive, and the students that are not responsive may be candidates for special education.

A core component of MTSS is data-based decision making—that all decisions are based on assessment data. Data are collected on all students through a screening process. Screening takes place at regular intervals through the school year by measuring performance on skill-based and behavioral indicators found through research to predict future academic success (Glover, 2010a). Data are then analyzed to determine whether the majority of students meet benchmark expectations in response to core instruction and which students may require additional interventions. Progress is closely monitored at each tier of intervention to determine the need for progressively intense instruction (Hughes & Dexter, 2011). Progress monitoring should be: (a) conducted frequently; (b) a

quick and easy method for gathering data; (c) able to analyze student progress to modify instruction; and (d) informative enough to allow adjustment of student goals (Morris, 2013). Clearly defined data-based rules are used to make decisions about students' eligibility for specific instructional programs and interventions (Glover, 2010a).

There is a lack of research on how to frame writing within an RTI model (Saddler & Asaro-Saddler, 2013). Hughes and Dexter (2011) examined 13 field studies to examine RTI efficacy. The outcomes of these studies relate primarily to early reading and math skills. Few, if any of the studies examined the impact of RTI on higher-level reading or math skills, writing, or in content areas such as science or social studies. To determine how writing can be addressed within the RTI framework, it is necessary to review how writing is currently assessed in schools.

Writing Assessment

Writing assessment has long been considered a problematic area (Huot, 1990). Diederich, French, and Carlton (1961) conducted a large-scale study evaluating interrater reliability with 300 essays read by 53 judges. They found that 94% of the essays received at least seven different scores. As early as 1912, essay scoring was proclaimed problematic because it was unreliable (Starch & Elliot, 1912). The alternative was to test students' writing ability indirectly with examinations consisting of multiple choice items covering grammar, usage, and mechanics (Huot, 2002). Until the mid-1960s, this was considered the only reliable and accurate way to evaluate students' writing ability (Huot, 1990).

To determine which writing measure to use, the following questions may be asked: (1) What is the purpose of the assessment? (2) What information is needed from the assessment? (3) How will the assessment information be used? (4) Which assessment

procedures fit the needs (Penner-Williams et al., 2009)? The Standards of the Educational and Psychological Testing (American Educational Research Association, 2014) stated that "…justification for the role of each instrument in selection, diagnosis, classification, and decision making should be arrived at before test administration, not afterwards" (p. 114).

There has also been a lack of alignment between writing theories and the practice of writing assessment. The gap between theory and assessment practices seems to be widening, especially in the area of high-stakes assessments. According to Behizadeh and Engelhard Jr. (2011), the underlying theory of these high-stakes assessments align more with writing as a skill than writing to produce meaning within a social context. Most writing assessments have been created by the measurement community with little input from writing theorists or individuals in the English teaching profession because of the overriding concern of reliability and validity (Huot, 2002).

Assessment quality. Reliability and validity are two of the most basic concepts in measurement theory (Cherry & Meyer, 2009; Pierangelo & Giuliani, 2012; Thorndike, 1990). In the area of writing, measures must yield consistent results (reliability) and must actually measure writing ability (validity). These two components of assessment quality are also related, as reliability is a necessary, but not sufficient component of validity. Therefore, it is important to have procedures to determine both validity and reliability of writing assessments.

Validity is the extent to which a test measures what it claims to measure. It is vital for a test to be valid in order for the results to be accurately applied and interpreted (Cherry & Meyer, 2009). Validity is not measured by a statistical procedure or test; it is

determined by a body of research that describes the relation between the assessment and the behavior it is measuring. The three sources of evidence for any test's validity include its content, its relation to the underlying "construct," and its ability to predict scores on related "criterion" measures, otherwise known as content validity, criterion validity, and construct validity (Thorndike & Thorndike-Christ, 2010).

Content validity depends on the extent to which a test reflects a specific domain of content (Thorndike & Thorndike-Christ, 2010). Content validity is usually measured by relying on the knowledge of people who are familiar with the construct being measured, known as subject-matter experts. The Common Core State Standards have recently been used as the content domain comparison. In essay scoring, content validity also refers to the rubrics that are used to assess the piece of writing. It must be determined if the criteria on the rubric match the knowledge and skills of the writing domain.

Criterion validity refers to the extent to which one measure estimates or predicts the values of another measure (Eaves & Woods-Groves, 2007). The first measure is known as the predictor variable. The second measure is called the criterion variable. The criterion variable usually has established validity, so the evaluators are seeking a strong relation between these two measures. There are two different types of criterion validity: concurrent validity and predictive validity. Concurrent validity occurs when the two measures are obtained at essentially the same time (Thorndike & Thorndike-Christ, 2010). This indicates the extent to which the test scores accurately estimate an individual's current state with regards to the criterion. Predictive validity follows the same procedures, but the criterion variable is administered at a later date. Regression

analysis is often used to determine criterion validity. The correlation coefficient between these two variables is referred to as the validity coefficient.

Construct validity is the degree to which a test score measures the psychological or cognitive construct that the test is intended to measure. To determine the construct validity of a writing measure, evaluators identify the factors that contribute to individual's performance on the measure, with evidence emerging from the conceptual framework. Construct validity can be measured by using content analysis, correlation coefficients, factor analysis, ANOVA studies demonstrating differences between differential groups or pretest-posttest intervention studies, and factor analysis (Brown, 1996).

Reliability is the other form of assessment quality that needs to be evaluated when analyzing measures. Reliability is an estimate of a measure's accuracy and consistency, otherwise known as an estimate of the extent to which a score measures the behavior being assessed (Greenberg, 1992). No score is perfectly reliable; some degree of inconsistency is present in all measurement procedures (Thorndike & Thorndike-Christ, 2010). Sources of error in any measurement situation include inconsistencies in the behavior of the participants, variability in the administration of the measure, and differences in raters' scoring behaviors. Test developers strive to reduce these impacts to produce reliable measures (Pierangelo & Giuliani, 2012). There are four different methods to assess reliability: (1) test-retest reliability; (2) split-half reliability or internal consistency; (3) interrater reliability; and (4) alternate forms reliability.

Interrater reliability is typically used to determine reliability on essay writing measures (Greenberg, 1992). Interrater reliability is normally done by administering the

writing prompt and having an objective scorer also score the same sample (Overton, 2009). The results are then correlated to determine how much variability exists between the scores. This procedure is especially important when measures have a great deal of subjectivity (Pierangelo & Giuliani, 2012). Test-retest reliability suggests that the participants will obtain a similar score when tested at a different time. In general, the shorter the test –retest interval, the higher the reliability coefficient (Sattler, 2008). This method to determine reliability is also used when evaluating different essay writing assessments and scoring techniques (Greenberg, 1992).

The other two forms of reliability are often used in norm-referenced assessments (Pierangelo & Giuliani, 2012). Alternate forms reliability is used when instruments have two or more different forms. The equivalent tests are given to the same group of examinees and correlations are determined by comparing these individual scores (Cohen & Spenciner, 2011). Internal consistency refers to the degree to which a student's answers to items measuring the same trait are consistent. Split-half reliability is commonly used to determine internal consistency (Pierangelo & Giuliani, 2012). It is determined by computing each individual's total score on even numbered items and correlating it with the odd scores.

Both instrument reliability and criterion-related validity are context-bound (Cherry & Meyer, 2009). Instrument reliability cannot be generalized to assessment situations that do not correspond to the original one in terms of the students, the test itself, and the assessment procedures. Criterion-related validity is also similarly limited. The predictive validity of a test, for example, is not better than the validity of the tests with which it correlates (Pierangelo & Giuliani, 2012). Another concern relates to how

reliability and validity coexist. The history of writing assessment shows that achieving high reliability in writing is not easy, but researchers need to make sure that validity is not sacrificed to achieve high levels of reliability (Moss, 1994; Rezaei & Lovorn, 2010). Strongly developed scoring approaches are commonly used to reduce these concerns related to reliability.

Scoring Approaches

There are two major approaches used to directly assess students' writing performance: quality and quantity measures. These approaches have been developed to assess writing performance. Both approaches use specific methods to analyze students' writing samples. Since their inception, researchers are continually studying methods address concerns related to reliability and validity that accompany writing assessments.

Quality measures. Quality measures use specific methods to determine the overall quality of the writing. Characteristics of quality writing are determined, and then different methods are used to evaluate pieces of writing based off of those characteristics. These approaches are used to assess writing quality at both large-scale and individual class level (Gearhart, 1992; Graham, Harris, & Hebert, 2011; Huot, 1990). The methods include holistic, analytic, and trait scoring. In holistic scoring, a single rating of general quality of the composition is made (Graham et al., 2011). Analytic scoring uses rubrics that allow the raters to evaluate one characteristic at a time (Miller, Linn, & Gronlund, 2013). Trait scoring is used to analyze traits specific to the writing product (Kim, Al Otaiba, Folsom, Greulich, & Puranik, 2014). All of these methods are currently used to assess writing in different program and products.

Holistic scoring. Early in its development, holistic scoring seemed to be overly subjective and unreliable (Huot, 1990). With a focus on training and calibration of

scoring techniques, holistic scoring is now widely used for research, placement, and other situations involving large numbers of writers (Calfee & Miller, 2007). This method is popular because of its unique combination of validity, speed, and reliability (Holt, 1993). This allows scorers to assess a number of samples quickly with a relatively strong level of both reliability and validity. It also generally takes less time than the other scoring procedures (Miller et al., 2013). As a result, holistic scoring dominates large-scale assessments (Calfee & Miller, 2007). For example NAEP, the only national writing assessment, uses a 6-point holistic scoring rubric (Miller et al., 2013). Based on a recent survey of state writing assessments, 67% of the states with writing assessments assign holistic scores (National Writing Project, 2008). To be able to implement this scoring procedure in such large scales, it is important that common methods are used.

Holistic scoring procedures usually follow similar procedures. Raters undergo intense training. The training includes reviewing anchor papers that are prototypes for each of the score categories. The rater gives the composition a brief reading, usually only one or two minutes, and assigns it a single score. Benchmark papers are also inserted during the scoring process to allow raters to recalibrate as needed. A detailed rating scale is often used that specifies multiple proficiency levels where each criterion is defined by specific descriptors (Hamp-Lyons, 1991). There are typically 4 to 7 proficiency levels. This process leads to reasonable high interrater reliability, a designation of rater agreement (Calfee & Miller, 2007). Burgin and Hughes (2009) suggested that teachers work in rater pairs and that each sample should be scored twice. These authors suggest that teachers also use a common rubric throughout the year to increase reliability.

Analytic scoring. Analytic scoring focuses on several identifiable qualities associated with good writing, and the rubrics are designed around these familiar writing features (Calfee & Miller, 2007; Huot, 1990). Analytic scales produce separate ratings for specific attributes such as ideation, organization, and style (Graham et al., 2011). These rubrics enable a teacher to focus on one characteristic of a response at a time (Miller et al., 2013). For example, the isolation of writing mechanics from the quality of the content might be especially useful when noting specific strengths and weaknesses (Gunning, 2002) These separate scores enable teachers to give clear and focused feedback (Miller et al., 2013). Oregon uses an analytic scoring rubric for its statewide writing assessment. The rubric consists of seven analytic dimensions: ideas and content; organization; voice; word choice; sentence fluency; conventions; and citing sources when required (Miller et al., 2013).

Since they were first developed, many analytic scales have been used to identify components of writing quality (Diederich, 1974), but scoring procedures to use analytic rubrics have remained consistent. Raters give scores to these individual qualities, and the scores are tallied to provide an overall rating of the writing sample (Huot, 1990). Rubrics used with young writers are usually simple, focusing on discernable factors. A rubric for a first or second grade student might focus on indentation, capitalization, punctuation, high-frequency words spelled correctly, and credible attempts at lower-frequency words (Calfee & Miller, 2007). As students progress in grades, rubrics become quite detailed, focusing more on the content. A fifth grade rubric might contain elements related to opening, middle, and closing paragraphs, voice, content reflecting research, and use of the author's own ideas (Calfee & Miller, 2007). Miller (1995) recommends using a rubric

that follows the five elements of writing: clarity; support of main ideas and subpoints; organization and development; mechanics, including grammar, spelling, punctuation, capitalization, paragraphing, and sentence structure; and overall rating addressing how the student performed the specific writing task.

Trait scoring. Primary trait scoring is a more specific form of analytic rubric scoring. It involves the identification of one or more traits relevant to a specific writing task (Huot, 1990). These traits are related to specific rhetorical situations created by the purpose, audience, and writing assignment. The most common trait scoring system is Six Trait or the 6+1 trait system (Spandel, 2013). The traits include: (a) ideas for how well main ideas were developed and represented; (2) organization for text structure; (3) word choice for use of interesting and specific words; (4) sentence fluency for grammatical use of sentences and flow of sentences (5) spelling for accuracy and for the developmental phase of spelled words; (6) mechanics for capitalization and punctuation accuracy; and (7) handwriting for spacing, neatness, and letter formation. 6+1 trait scoring is a popular system used in schools, but there is little empirical evidence about the factor structures, with only two studies in the literature (Kim et al., 2014).

Technical adequacy of quality measures. Numerous studies have analyzed the differences between scores obtained using both holistic and analytic scoring methods (Barkaoui, 2011; Harsch & Martin, 2013). These studies found that holistic scoring led to higher levels of rater agreement. Even though they were more consistent, Barkaoui (2011) stated that raters were harsher when using holistic scoring methods. This might be because some students have different levels of proficiency in different areas of writing, or an uneven profile. This is difficult to capture in a single holistic score; therefore, the

holistic score only captures the low skill levels. Holistic scores may also mask deviances in how descriptors are applied (Harsch & Martin, 2013). With a holistic score, it is not possible to examine how raters applied the different parts of the scale in order to form their holistic criterion scores or overall scores. To address these concerns, analytic descriptors and rubrics could be used alongside holistic methods to accurately assess students' writing samples.

Graham and colleagues (2011) conducted a meta-analysis of scoring procedures used in writing assessments with both holistic and analytic scoring procedures. They analyzed 22 studies that examined the reliability of holistic scales on everything from high-stakes writing assessments to more typical classroom assessments, including portfolios. In each of these studies there were more than two raters, increasing the generalizability of the findings. To evaluate reliability, the authors used two different approaches, consensus and consistency. The Consensus Approach calculates the percentage of exact agreement to indicate how often raters assign the exact same score. The Consistency Approach calculates a reliability coefficient to provide an estimate of the degree to which the pattern of high and low scores among raters is similar. An exact percentage of agreement of 70 percent or better indicates reliable scoring with the consensus approach (Brown, Glasswell, & Harland, 2004). A reliability coefficient of .80 is generally viewed as acceptable with the consistency approach (Nunnally, 1967; Shavelson & Webb, 1991), but higher coefficients are desirable when scores are used to make decisions about individual students (e.g., a reliability coefficient between .90 and .95). Only 25% of studies met the 70% criteria for consensus, whereas 47% of the studies met the .80 criteria for consistency. Twenty-one similar studies that examined the

reliability of analytic scales were found. None of the studies met the 70% criteria for consensus; only 26% of the studies met the .80 criteria for consistency. Graham and colleagues (2011) stated that care must be given to establish the reliability of subjective writing measures, such as holistic and analytic writing scales, if teachers are to use these scoring procedures in their classrooms. Otherwise, scores from these measures will be unreliable for teachers to make sound decisions about students' writing on their progress as writers.

Some researchers believe that well-constructed rubrics and scoring procedures may promote stronger reliability scores (Burgin & Hughes, 2009; Gearhart, 1992; Gunning, 2002; Jonsson & Svingby, 2007). Rubrics are used for many purposes and different age groups, from early childhood to higher education (Humphry & Heldsinger, 2014). Even though analytic rubrics have emerged as one of the most popular tools in progressive education programs, there is a lack of empirical evidence in the literature quantifying the actual effectiveness of rubrics as an assessment tool (Andrade, Du, & Mycek, 2010; Humphry & Heldsinger, 2014; Rezaei & Lovorn, 2010).

Concerns exist related to reliability and validity of rubrics (Humphry & Heldsinger, 2014). Studying rubric reliability has found low reliability coefficients when compared to traditional psychometric requirements (Jonsson & Svingby, 2007. This indicates that the use of a rubric might not in itself be enough to produce sufficient reliability. This may not be an indication of fault with rubrics per se, but with the users. A large-scale study attempted to see if a rubric could increase reliability when compared to holistic scoring (Rezaei & Lovorn, 2010). The researchers gave two different papers to college students that served as the raters. One paper was well written in terms of skills

and mechanics, but only addressed a broad description and did not answer any of the elements of the prompt. The paper also did not contain any citations, which was a requirement of the prompt. The second paper addressed all of the parts of the prompt with a variety of sources. This paper did have multiple mechanical errors, but writing mechanics only constituted 10% of the rubric. According to the rubric, this paper deserved a high score, one much higher than the first paper. Instead, the opposite occurred. The second paper scored lower on "understanding and synthesis of argument", "understanding the goals and implications of the topic" and "support and citation of sources" even though the first paper did not address these components. The use of a rubric also did not lessen the range of assigned scores. Although this rubric was designed to reduce or eliminate raters' bias, writing mechanics was still obviously a significant factor in raters' assessments. This was especially concerning to the authors in regards to students learning English and the mechanical rules of the language. Not only did this study show low reliability of the raters, it also questions the rubric's construct and criterion-based validity. To increase specification and reduce the amount of inconsistencies between raters, primary trait scoring may be used.

Technical characteristics of CBM measures and Six Trait measures were compared using the Stanford-9 as the criterion measure (Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006). Reliability, as measured by interobserver agreement, was adequate, but exact agreement on the Six Trait measures was low. The authors contend that Six Trait measures do not measure distinct components of writing, nor do they share a significant amount of variance with Stanford-9 measures of written expression resulting

in small validity coefficients. The authors continue that the results of this study do not support the use of a trait model as a measure of written expression.

In the other study regarding 6+1 trait scoring, Kim et al. (2014) collected 531 writing samples from first graders to examine the dimensions of writing compositions. The samples were scored using 6+1 trait scoring and indicators for production and syntax. The 6+1 traits included: ideas; organization; word choice; sentence fluency; spelling; mechanics for capitalization and punctuation; and handwriting. These components were rated on a scale of 1 to 5. Productivity indicators included: number of words; number of different words; and number of different ideas. The syntactic complexity was determined by the mean length of the T-unit and clause density. T-units were defined as one independent clause plus any dependent clauses and clause density was calculated as a ratio of the total number of clauses divided by the total number of T-units. This study found that the written compositions were only composed of two dimensions: substantive quality and conventions including spelling. The traits ideas, organization, sentence fluency, and word choice shared enough common variance to be described as one single dimension. The findings of this study indicate that giving each trait equal weighting may reduce the validity of the scores.

Quantity scoring. Even though quantity scoring can be applied to any writing sample, the most common use of quantity scoring is in CBM. Developed by Deno and colleauges in the early 1980s, CBM gave teachers academic measures that could be collected daily, graphed, and evaluated for evidence of student learning within short periods of time (Shinn & Bamonto, 1998). There are many characteristics specific to CBM. Moving away from indirect forms of assessment that require students to answer

multiple choice questions, academic performance is sampled by direct observation procedures. Students may read, write, or complete mathematical computations while completing CBM measures. Scores are obtained by counting the number of correct and incorrect responses made in a fixed time period. The use of multiple equivalent samples is one of the most distinctive and important features of CBM. Students respond to different but equivalent stimulus materials that are drawn from the same general source.

Another characteristic of CBM is that the measures are designed for efficiency. CBM performance samples are 1-3 minutes in duration, depending on the skill being measured and the number of samples necessary to maximize reliability. Efficiency is important so that significant amounts of precious instructional time are not lost to testing (Shinn & Bamonto, 1998). Teachers need to make decisions about the effectiveness of instruction as quickly as possible (Hessler & Konrad, 2008). This allows teachers to quickly identify ineffective instruction and modify it as needed. The last characteristic identified by Deno (2003) referred to the ease with which professionals, paraprofessionals, and parents can learn to use the procedures in such a way that the data are reliable.

CBM measures have been thoroughly analyzed for implementation and use. First, the measures are technically adequate (Deno, 2003). The reliability and validity of CBM have been achieved through the use of standardized procedures for repeatedly sampling performance on core reading, writing, and arithmetic skills. To address reliability, specific directions are provided with scripts that are read before each administration. There are also specific directions related to timing and scoring to assist with reliability. Validation occurs by using correlation analysis between CBM measures and outcome

measures (Wallace, Espin, McMaster, Deno, & Foegen, 2007). Second, the measurement tasks, the behaviors expected from the student, are standardized. The standard tasks identified for use include reading aloud from text and selecting words deleted from text in reading, writing word sequences when given a story starter or picture in writing, writing letter sequences from dictation in spelling, and writing correct answers/digits in solving problems in arithmetic. This standardization ensures that changes in test scores are attributed primarily to student improvement rather than changes in testing conditions (Shinn & Bamonto, 1998). Third, specifications are provided related to material selection. Key factors in this selection process are the representativeness and equivalence of the stimulus materials. Both factors are addressed to increase the utility of the procedures for making instructional decisions.

CBM can be used to: (a) improve individual instructional programs; (b) predict performance on important criteria; (c) enhance teacher instructional planning; (d) develop norms; (e) increase ease of communication; (f) screen students to identify those academically at risk; and (g) evaluate classroom prereferral interventions. The systematic approach to setting goals, monitoring growth, changing programs, and evaluating effects of changes allows this system to improve individual instructional programs. CBM data have been used to predict future success or difficulty on a number of outcome measures. It can be used not only to evaluate instruction, but also predict and improve on teacher judgments regarding student proficiency, identify students' discriminate between students achieving typically and those in compensatory programs, and predict who will succeed on high-stakes tests (Deno, 2003; Fuchs, 2004; Roberts et al., 2012; Wallace et al., 2007). CBM can also be used to enhance teachers' instructional planning by assisting teachers in

identifying their students' strengths and needs (Deno, 2003; Hessler & Konrad, 2008). CBM measures have also been used to facilitate communication between teachers, students, and families. It is common practice for teachers to use the CBM data in parent conferences and at multidisciplinary team meetings to provide a framework for communicating individual student status (Deno, 2003; Hessler & Konrad, 2008).

Even though CBM has been historically used by special educators to monitor students' performance, it has more recently been used in general education community for screening decisions. Not only does it allow teachers to identify those that are at risk academically, it also allows teachers to identify students that will be successful in school (Deno, 2003; Roberts et al., 2012). This assessment system allows teachers to track progress across the first few years of school (Roberts et al., 2012). Based on this information, teachers are then able to provide individualized instruction to all students. This will include early intervention to those students at risk for academic failure. It is important to recognize that CBM is an indicator, meaning that research has shown that CBM measures correlate with key behaviors indicative of overall performance in the academic areas of reading, math computation, written expression, and spelling (Shinn & Bamonto, 1998). As an indicator, it does not sample all behaviors within an academic domain. CBM data should not be the only assessment data collected in a particular domain.

CBM data are often collected to measure progress throughout the intervention to determine effectiveness because measures are sensitive to the effects of program changes over relatively short time periods (Deno, 2003). This also provides teachers with the data they need to document student progress if more intensive interventions or special

education is necessary (Hessler & Konrad, 2008). Based on the need for valid and reliable measures for screening and progress monitoring within the RTI framework, Fuchs (2004) outlined three stages of research needed to establish the utility of CBM for these uses. The first stage involves examining the technical features of the static score. The static score refers to the performance level at one point in time. Studies at Stage 1 typically test a measure's validity by correlating the scores to an outcome measure (Glover, 2010b). Reliability is also often tested at Stage 1. Stage 2 involves examining technical features of the slope. Since each weekly test is comparable in difficulty and conceptualization, slopes can be used to quantify rate of learning (Fuchs, 2004). Stage 3 then examines the instructional utility of the measure. Most CBM-W research is still in Stage 1, focused on the reliability and validity of static scores. Little focus has been placed on screening, monitoring progress, and designing and evaluating effective writing interventions (McAlenney & McCabe, 2012). One overriding reason for this is that additional research is still needed to identify a writing measure that can accurately identify students with writing difficulties (Saddler & Asaro-Saddler, 2013).

CBM-W typically involves giving students a writing prompt with specified time to plan and write. Countable indices are then used to score the sample (De La Paz, 2007). Writing fluency refers to the natural flow and organization of a piece of writing. Fluent prose is easier and more enjoyable to read as the words are organized in a logical fashion and the overall message is easy to understand (Saddler & Asaro-Saddler, 2013). Simple fluency, or production dependent, measures include: Total Words Written (TWW); Words Spelled Correctly (WSC); and Correct Letter Sequences (CLS). In addition, total punctuation marks, correct punctuation marks, number of sentences, and words in

complete sentences have also been used (Saddler & Asaro-Saddler, 2013). A composite of several production-dependent measures is Correct Word Sequence (CWS). Production-independent indices, or measures of accuracy, have also been used in a number of research studies. These measures include ratios of correct to incorrect observations of the indices used in production-dependent measures. For example, Percentage of Words Spelled Correct (%WSC) refers to the percentage of words spelled accurately from the entire passage. Even though research has been conducted on these different measures since 1980, no consensus exists as to which measures teachers should use or for whom these measures work best (Saddler & Asaro-Saddler, 2013).

Technical adequacy of CBM-W. Stanley Deno and his colleagues began studying CBM in the area of writing along with his original CBM work with reading through the Institute for Research on Learning Disabilities (IRLD) at the University of Minnesota starting in 1980 (McMaster & Espin, 2007). One of those first studies assessed students in grades 1-6 on WSC, CLS, and TWW at fall, winter, and spring (Marston, Lowry, Deno, & Mirkin, 1981). On WSC, results indicated fairly steep incremental growth. CLS and TWW did not show the same growth pattern, but growth did occur from fall to spring as expected. Based on the immediate and dramatic growth seen at the early grade levels, the authors stated that these measures were sensitive enough to be useful for evaluating instructional programs for students that have learning disabilities even though no special education students were included in the sample.

The study conducted by Deno, Marston, and Mirkin (1982) provided an empirical base for developing validation procedures used in many current studies. These researchers examined correlations between TWW, WSC, CLS, large words, mean t-unit

length, and mature words with the Test of Written Language (TOWL), Stanford Achievement Test and Developmental Sentence Scoring System. TWW, WSC, CLS, and Mature Words most strongly and consistently related to the criteria used. Deno and colleagues stated that since TWW is the most time-efficient scoring procedure, TWW should be used.

CWS was first examined in a study with 50 students in 3rd-6th grade (Videen, Deno, & Marston, 1982). This study found a high degree of agreement on reliability measures using CWS. Scores also ranged from 27.3 for third grade to 58.8 for sixth grade, indicating sensitivity of the measure. CWS also correlated significantly with the holistic rating scale used, TWW, WSC, the raw total on the TOWL and Developmental Sentence Scoring. Even though the time needed to score CWS is substantially greater that TWW or WSC, it was found to be a valid and reliable measure of written expression that could be used in a formative evaluation system. A follow-up study analyzed CWS along with TWW, WSC, Words Spelled Incorrectly, and CLS. The results indicated CWS was the most discriminating measure between adjacent grade levels.

These early studies conducted by the IRLD focused on determining if production variables were valid and reliable. These studies used small sample sizes and often reported correlations across grades (McMaster & Espin, 2007). Subsequent research reported substantially less positive results when analyzing students within a specific grade level. In a study that evaluated five different criteria (TWW, number of words in a t-unit, number of different words used, ratio of different words to total number of words, and number of words with seven letters or more), only TWW produced significant

differentiation among ability groups or by grade level (Barenbaum, 1987). This continued to support TWW over other measures assessing production of words.

Additional research has explored which CBM measures were most appropriate to use with various groups of students. Eight different measures were used to score six-minute writing samples from 6th-8th graders (Tindal & Parker, 1989). The production variables (TWW, number of legible words, and WSC) were weakly correlated with the holistic ratings of communicative effectiveness scored by the teachers. However, two production-free factors, %WSC and %CWS, were highly related to the holistic scores. The variables did fail to discriminate between students in special education programs and their general education peers. This study aligned with future research that shows that production dependent variables are not as valid of an indicator for older writers (Amato & Watkins, 2011; Espin et al., 2005; Espin et al., 2000; Jewell & Malecki, 2005; Weissenburger & Espin, 2005).

Along with teacher rating scales, CBM measures were analyzed to identify direct writing measures that were technically adequate for large-scale program evaluations (Tindal & Parker, 1991). In this study, ten minute writing samples were obtained from 3rd-5th graders at the beginning and end of the school year. Measures included three production indices (TWW, WSC, and CWS) and three quality measures scored using an analytic scoring system (story idea, organization-cohesion, and convention-mechanics). The correlations between the measures were not uniformly high, especially between qualitative and quantitative measures. The authors did find significant group differences within grade levels between students with learning disabilities and typical students. Students with LD consistently wrote less, wrote less correctly, and were not well judged

on either content or organization. The authors assert that since students with LD write very little, a simple count of TWW may be sufficient. However, this study found that many students show little change on the production measures, but write better compositions when judged qualitatively after a year of instruction. These results support a multi-faceted effort in evaluating writing within any program evaluation.

Progress monitoring with CBM-W was first used in a study that investigated the technical adequacy of seven objective indices (Parker, Tindal, & Hasbrouk, 1991b). The participants included 36 middle school students with mild disabilities that were assessed four times over a six-month period. The CBM measures included: TWW; legible words; WSC; CWS; average length of all continuous strings of CWS; %total legible words and; %WSC. Holistic ratings of the same samples and TOWL writing samples were used as criterion measures. On the basis of direct assessment and informal judgments, students appeared not to improve in writing over the six months. However, there was pronounced linear growth for TWW, legible words, and WSC. TWW and legible words yielded the lowest correlations in static comparisons with holistic ratings and the TOWL assessment. The authors noted that even though the three objective indices suggest "sensitivity growth", the conclusion is not corroborated by the criterion measures or informal observations by the research team. This study provided further skepticism of using TWW and WSC with older writers. The indices that were most highly correlated with TOWL and the holistic ratings were CWS, mean lengths of CWS, and %legible words, providing promise for these indices to be used in screening and eligibility decisions.

A final study conducted by Parker, Tindal, and Hasbrouck (1991a) used six-minute writing samples with students in grades 2-6, 8, and 11 in fall and spring. Five

indices were analyzed: TWW, WSC, CWS, %WSC, and %CWS. Holistic ratings were used as validation. Using mean scores across grade levels and assessment periods, all five indices appeared suitable. When analyzed by individual grade levels, only %WSC was suitable for second grade. In third grade, CWS and %CWS were suitable with TWW designated as marginal. In fourth grade, %WSC, %CWS, and TWW were suitable and WSC was marginal. The authors also state that based on their findings, only the percentage-based indices proved to be generally suitable for screening and eligibility purposes, but standard error of measurement bands existed for all the indices.

After this study, little research examining CBM-W occurred for ten years. CWS-IWS was validated for use with middle school, further indicating the weakness of TWW and WSC for older writers (Espin et al., 2000). Another study analyzing middle school students reading and writing performance found that WSC on untimed samples helped detect significant differences in student performance, but are not complete measures of overall writing competence (Fewster & Macmillan, 2002). The validity levels for the writing measures were also much lower than the reading measures that were investigated.

In the early 2000s, research emerged at the elementary level examining new CBM measures to replace TWW and WSC, outlined as valid by Marston's review (1989) of the early CBM-W work. One study analyzed how third and fourth graders' CBM scores correlated to standardized tests and teacher ranking (Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002). Fourteen different measures were used, and CWS and correct punctuation marks were found to be promising indicators of written expression. TWW was not as strong as the other indicators. Total punctuation marks, simple sentences, and words in complete sentences emerged as the best predictors in a study of how CBM

measures compare to the Woodcock Johnson-Revised (WJ-R) writing samples subtest with third and fourth grade students (Gansle et al., 2004). TWW did not enter the regression equation that predicted the WJ-R, but did respond to a brief intervention that was given during the study. The other metrics did not respond to the intervention. This study reiterates finding from Parker et al. (1991a) that TWW is sensitive to growth, but not as valid as other indicators.

Six different indices were examined as indicators for national and state standardized tests and grades in language (Jewell & Malecki, 2005). The participants were students were in 2nd, 4th, and 6th grade. Boys were found to score lower than girls on the fluency measures (TWW, WSC, and CWS) but statistical differences were not found on the accuracy indictors (%WSC, %CWS, and CWS-IWS). Overall, measures of accuracy more strongly related to students' performance on other types of writing than fluency measures, with all correlations decreasing with the increase in grade level. This line of study was continued by assessing students in grades 4, 8, and 10 to determine if CBM measures differ for students as they get older (Weissenburger & Espin, 2005). A statewide achievement test was used as the criterion validity measure. The CBM measures included TWW, CWS, and CWS-IWS. Criterion-related coefficients were stronger for CWS and CWS-IWS than TWW at all grade levels, with no differences of CWS and CWS-IWS between fourth and eighth graders. Even though CWS and CWS-IWS was found to be a valid indicator for fourth and eighth graders, no CBM was found to be a good indicator of general writing proficiency for high school students. Espin et al. (2005) also found a strong relation between CWS and CWS-IWS when correlated with qualitative ratings of expository essays. Amato and Watkins (2011) also conducted a

study with 8th grade students that provided limited support for CBM-W indices. This study analyzed seven different metrics on untimed writing samples. Together, the indices accounted for only 44% of the variance in TOWL-3, with three of the indices (CWS, correct punctuation marks, and correct capitalization) uniquely contributing to the prediction of TOWL-3 scores.

Table 1

Summary of Key CBM-W Concurrent Validity Studies

CBM-W Metric	Validity Coefficient	Grade	Criterion Measure	Study
TWW	.65	3-6	TOWL- WLQ	Deno et al., 1982a
	.20	3	Analytic Scoring	Tindal & Parker, 1991
	.42	4	Analytic Scoring	Tindal & Parker, 1991
	.37	5	Analytic Scoring	Tindal & Parker, 1991
	.49	2	Holistic Rating	Parker et al., 1991b
	.40	3	Holistic Rating	Parker et al., 1991b
	.36	4	Holistic Rating	Parker et al., 1991b
	.44	5	Holistic Rating	Parker et al., 1991b
	.15	3	ITBS	Gansle, 2002
	.23	3-4	WJ-R Writing	Gansle, 2004
	.24	2	SAT Language	Jewell & Malecki, 2005
	.22	4	SAT Language	Jewell & Malecki, 2005
	-.14	6	SAT Language	Jewell & Malecki, 2005
	.48	4	WKCE Writing	Weissenburger & Espin, 2005
CWS	.69	3-6	TOWL- Raw Score	Videen et al., 1982
	.85	3-6	Holistic Rating	Videen et al., 1982
	.29	3	Analytic Scoring	Tindal & Parker, 1991
	.48	4	Analytic Scoring	Tindal & Parker, 1991
	.34	5	Analytic Scoring	Tindal & Parker, 1991
	.60	2	Holistic Rating	Parker et al., 1991b
	.58	3	Holistic Rating	Parker et al., 1991b
	.58	4	Holistic Rating	Parker et al., 1991b
	.61	5	Holistic Rating	Parker et al., 1991b
	.43	3	ITBS	Gansle, 2002
	.36	3-4	WJ-R Writing	Gansle, 2004
	.57	2	SAT Language	Jewell & Malecki, 2005

Table 1- Continued

	.46	4	SAT Language	Jewell & Malecki, 2005
	.23	6	SAT Language	Jewell & Malecki, 2005
	.59	4	WKCE Writing	Weissenburger & Espin, 2005
WSC	.67	3-6	TOWL- WLQ	Deno et al., 1982a
	.31	3	Analytic Scoring	Tindal & Parker, 1991
	.45	4	Analytic Scoring	Tindal & Parker, 1991
	.36	5	Analytic Scoring	Tindal & Parker, 1991
	.64	2	Holistic Rating	Parker et al., 1991b
	.54	3	Holistic Rating	Parker et al., 1991b
	.49	4	Holistic Rating	Parker et al., 1991b
	.56	5	Holistic Rating	Parker et al., 1991b
	.24	3	ITBS	Gansle, 2002
	.38	2	SAT Language	Jewell & Malecki, 2005
	.29	4	SAT Language	Jewell & Malecki, 2005
	-.05	6	SAT Language	Jewell & Malecki, 2005
%CWS	.43	2	Holistic Rating	Parker et al., 1991b
	.53	3	Holistic Rating	Parker et al., 1991b
	.70	4	Holistic Rating	Parker et al., 1991b
	.55	5	Holistic Rating	Parker et al., 1991b
	.59	2	SAT Language	Jewell & Malecki, 2005
	.67	4	SAT Language	Jewell & Malecki, 2005
	.52	6	SAT Language	Jewell & Malecki, 2005
%WSC	.48	2	Holistic Rating	Parker et al., 1991b
	.49	3	Holistic Rating	Parker et al., 1991b
	.67	4	Holistic Rating	Parker et al., 1991b
	.55	5	Holistic Rating	Parker et al., 1991b
	.46	2	SAT Language	Jewell & Malecki, 2005
	.50	4	SAT Language	Jewell & Malecki, 2005
	.47	6	SAT Language	Jewell & Malecki, 2005

Note: TWW= Total Words Written; CWS= Correct Word Sequence; WSC= Words Spelled Correctly; TOWL= Test of Written Language; WLQ= Written Language Quotient; ITBS= Iowa Tests of Basic Skills; WJ-R= Woodcock Johnson-Revised; SAT= Stanford Achievement Test; WKCE= Wisconsin Knowledge and Concepts Examinations

Research analyzing the concurrent validity of CBM-W metrics has been conducted at grade levels from grades one through eleven with a variety criterion measures. Focusing on just elementary grades (see Table 1), TWW, CWS, WSC, %CWS, and %WSC were evaluated in numerous studies with mixed results. TWW is the most

inconsistent metric, with weak validity coefficients in recent research. CWS and %CWS continues to appear to be the strongest metric across all grade levels, with WSC and %WSC also providing strong correlations at the second grade level.

Recent research (Costa et al., 2012; McMaster et al., 2011; Parker, McMaster, Medhanie, & Silberglitt, 2011; Ritchey & Coker, 2013) has focused on the use of CBM-W for progress monitoring. This involves analyzing the slope and stability of the scores. Results from these studies did not conclusively support using the CBM-W measures that were analyzed for progress monitoring. Two studies that examined second grade students' scores found that slope estimates were relatively small and did not indicate sensitivity to change (Costa et al., 2012; Ritchey & Coker, 2013). The results from these studies are not surprising. Stage 1 research identifying CBM-W measures on static scores have not produced strong results; therefore, using those same measures in progress monitoring will not increase the validity of them. Before we can identify a measure to use in RTI to progress monitor writing progress, an initial measure for screening and identification must be developed. This review of literature indicates that new measures need to be analyzed as alternatives to what is currently being used in our schools. CBM-W and quality scoring methods both show promise, but neither indicates the level of validity needed to accurately screen and identify students in need of writing interventions.

CHAPTER THREE: METHOD

To determine effectiveness of writing interventions, it is necessary to have reliable and valid measures (Fuchs, 2004). In general, the psychometric properties of Curriculum-Based Measurement in Writing (CBM-W) indices are weaker than CBM measures in reading and math. The validity coefficients between CBM writing indices and traditional measures of writing may have lower values because of the complexities of measuring written expression skills. Some writing indices have shown potential for screening and progress monitoring as evident by the reliability and validity evidence, but additional indices might be needed. Shriner and Thurlow (2012) suggest adding a teacher quality measure to increase the predictive power of CBM-W for students' overall writing achievement. This study will evaluate the effectiveness of this procedure by answering the following research questions:

1. Which CBM metrics (TWW, words spelled correctly [WSC], CWS, percentage of words spelled correctly [%WSC] and percentage of correct word sequences [%CWS]), alone or in combination, best predict writing performance as measured by the TOWL-3?
2. What is the relation between the district writing assessment and validated writing measures?
3. What impact does adding an analytic rubric score generated from a classroom writing activity have on the CBM metrics' prediction of TOWL-3 performance?

Participants and Setting

This study took place in a rural school district in Eastern Iowa. It is a merged community school district serving approximately 1600 students in a 90-square mile area. The district is comprised of two incorporated towns with populations of approximately 2,400 and 1,400. Many residents within the district reside on acreages, farms, or in rural residences. The district currently has four buildings with a primary school (Prekindergarten-2nd grade), an intermediate school (3rd- 5th grade), a middle school (6th - 8th grade) and a high school (9th -12th grade). Almost 95% of the population is white and approximately 14% of the students qualify for free/reduced price lunches. Participants in this study were approximately 110 students in second grade, ranging from 7 to 8 years old. Currently, 12% of these students qualify for Special Education services. The students were placed into six different second grade classrooms. The teachers were all veteran teachers, ranging from 10-37 years of experience.

A Human Subjects Research Determination request was submitted to the Internal Review Board to determine if this study needed the committee's approval. Since the study does not meet the regulatory definition of human subjects, it was exempt. See Appendix A for a copy of the letter.

Measures

The primary school recently developed a district writing assessment with analytic rubrics. The principal of the school sought advice as to how to evaluate the validity of this assessment. It was suggested to use a validated writing assessment and compare the students' results on both assessments using correlation analysis. Without a state test to be used as a comparison, The Test of Written Language- 3 (TOWL-3) was used as a criterion measure to determine the validity of the district writing assessment. The TOWL-

3 is a test that has been validated for second grade and can be administered as a group (Hammill & Larsen, 1996). The current version (TOWL-4; Hammill & Larsen, 2009) could not be used because administration is not recommended until fourth grade. Since the TOWL-3 uses national norms, the administration also wanted local comparisons. The Area Education Association has normative information from CBM-W measures. Data from the 3-minute timed assessments was be used to further validate the district writing assessment through correlational analysis.

District writing assessment. Data from the district writing assessment were analyzed as a predictor variable. The students were given the following prompt:

> Think of the school experience we have at school: caring, positive attitude, respectful and responsible. Write a paragraph for you teacher that explains how a student follows two of the expectations.

The assessment allows students to use available planning tools. The students are not expected to edit or revise the sample, but no specific time limit is used. The teachers then evaluate each student's writing skills using an analytic rubric. The rubric (see Appendix B) includes six indicators: (a) content/ideas: (b) organization; (c) word choice; (d) conventions; (e) spelling; and (f) presentation. The first three components are directly related to the content of the composition. While assessing the first component, content/ideas, the raters determine if the writer responded to all parts of the prompt. The raters are also evaluating if the writers' topic and ideas are clearly communicated through the use of details or facts. Last, the raters evaluate if the questions of who, what, when, how, where, and why are sufficiently answered. The second component, organization, specifically focuses on sentence structure and beginning, middle, and end. The raters

determine if there is a clear introduction, body, and conclusion with supporting details. The next component addresses the students' word choice throughout the writing sample. In this section, word choice refers to the use of adjectives to accurately provide descriptions or "paints a picture for the reader." Nouns and verbs are also evaluated, along with the use of words to show feelings. This section also evaluates the sentence fluency or flow of the sample from one sentence to another with a variety of sentence constructions. The last three components of the rubric refer to accuracy in writing mechanics and presentation that are expected at this grade level. While assessing conventions, the raters are analyze correct conventions in capitalization and punctuation that are expected for second grade students. The expectation is that end punctuation marks are present and accurate, while punctuation within the sentences (commas and apostrophes) are mostly correct. Capitalization is expected at the beginning of the sentence and with the pronoun I. Correct capitalization within the sentence is expected most of the time. Spelling, the fifth component, is also evaluated at the developmental level of second graders. Accurate spelling is expected on the sight words and word patterns that have been explicitly taught, with accurate phonetic spelling for untaught words. In the last component, presentation, the rater answers the following questions with either yes or no: (a) uses correct spacing; (b) includes name and date; (c) letters are formed correctly; and (d) neat handwriting.

Each standard is rated based on a 3-point rubric (1 equals approaching standard; 2 equals developing standard; 3 equals at standard). A 0 can also be given if a student did not respond. Even though all five areas of the rubric are assessed, only the first three components related to the quality of the writing content are used to determine if a student

is at standard. The student must receive the full nine points on the rubric to be considered at standard. Since this a teacher-developed analytic rubric, there is currently not any reliability or validity evidence.

Table 2

CBM-W Scoring Metrics, Definitions, and Examples

CBM-W Metric		Definition	Example
Total Words Written	TWW	Number of total words written within the specified time limit; spelling, grammar, and content are not taken into consideration.	The tall boy sat down. TWW=5 The boy tall sat doun. TWW=5
Words Spelled Correctly	WSC	Number of total words written within the specified time limit; spelling, grammar, and content are not taken into consideration.	The tal boy sat doun. WSC=3 The boy tall sat doun. WSC=4
Correct Word Sequence	CWS	Number of adjacent, correctly spelled words that were syntactically and semantically appropriate given the context of the sentence; sequences are examined for correct meanings, tenses, number agreement (singular or plural), and noun-verb correspondences; punctuation, capitalization, and spelling are taken into account.	^The tal boy^sat doun. CWS=2 ^The tal boy sitted doun. CWS=1
%Words Spelled Correctly	%WSC	Ratio of the number of words spelled correctly to the total number of words written in the composition; formula is WSC/TWW x 100	The tal boy sat doun. %WSC=3/5=.60 The boy tall sat doun. %WSC=4/5=.80
%Correct Word Sequence	%CWS	Ratio of the number of correct word sequences to the total number of possible word sequences	^The--tal--boy^sat--doun.-- %CWS= 2/6=.34 ^The--tal--boy--sitted--doun.-- %CWS=1/6=.17

Curriculum-based measures of written expression. The curriculum-based writing metrics of total words written (TWW), words spelled correctly (WSC), correct word sequences (CWS), percentage of words spelled correctly (%WSC) and percentage of correct word sequences (%CWS) were scored from the 3-minute writing sample (see review of validity and reliability in Chapter 2). Each of the metrics consists of different scoring procedures (see Table 2).

Total words written. This metric is a count of the number of words written in a writing sample during the specific time limit. A word is defined as any letter or group of letters separated by a space, regardless of spelling. Spelling, grammar, and content are not taken into consideration when counting the number of words. Numerical representations and symbols are not included in the total.

Words spelled correctly. This is defined as the number of correctly spelled words written by the student within the time limit. Each word counted as correct must be able to stand alone in the English language. Context and grammar are not taken into account. A word might not be spelled correctly for the context of the writing sample, but if the word is recognized as a correct spelling of a word, it is counted correct. This index is calculated by subtracting the number of words in the writing sample that are spelled incorrectly from the total words written.

Percentage of words spelled correctly. The percentage of words spelled correctly is the ratio of the number of words spelled correctly to the total number of words written in the composition. The formula is (WSC/TWW) x 100.

Number of correct word sequences. Correct word sequences (CWS) is defined as two adjacent, correctly spelled words that are syntactically and semantically appropriate

given the context of the sentence. Words are examined for correct meanings, tenses, number agreement (singular or plural), and noun-verb correspondences. In addition, punctuation, capitalization, and spelling are taken into account when scoring CWS.

Percentage of CWS. The percentage of CWS is the ratio of the number of correct word sequences to the total number of possible word sequences.

Test of Written Language-3 (TOWL-3). The TOWL-3 was used as the criterion measure in this study. The TOWL-3 was developed to (a) help identify students with writing difficulties, (b) diagnose strengths and weakness of students' writing performance, (c) measure student progress in writing, and (d) conduct research in writing. The TOWL-3 was normed on a representative sample of 2,217 students in Grades 2 through 12. The Spontaneous Writing section used in this study assesses a student's ability to write a complete and interesting story. Students are given a picture prompt and are provided 15 minutes to write a story to go with the picture. They are encouraged to plan before they start writing. Three different subtests are then used to evaluate the writing sample. The first subtest is contextual conventions. The items on this subtest consist of evaluating the use of capitalization, punctuation (period, quotation marks, apostrophes, etc.) and spelling. The second subtest, contextual language, focuses on grammar (run-on sentences, subject-verb agreement, a/an appropriateness) and complexity of writing (naming of objects in picture, vocabulary selection, correct spelling of three syllable words). The last subtest, story construction, addresses the content of the story. These items focus on plot, story sequence, and characters. Each item is scored on a scale with either 2 options (e.g., Fragmentary sentences: 0 equals yes or 1 equals no), 3 options (e.g., Story beginning: 0 equals none, abrupt; 1 equals weak, ordinary,

serviceable; 2 equals interesting, grabbing), or 4 options (e.g., Introductory phrases or clauses: 0 equals none; 1 equals 1-2; 2 equals 3-5; 3 equals more than 5). These three subtests derive the Spontaneous Writing Quotient. This transformed score has a mean of 100 and a standard deviation of 15.

Reliability. To assess quality, tests are evaluated for reliability and validity. Reliability refers to the consistency of measurements across forms, different administration times, and different evaluators. As a consequence, if a test has poor reliability, the scores it produces are not stable, reproducible, predictable, dependable, meaningful, or accurate (Pierangelo & Giuliani, 2012). Tests with poor reliability will produce distinctly different scores when given at different times or when administered and scored by different people. Reliability is determined by estimating the degree of error associated with a test's scores (McMillan, 2012). If a test has high reliability, there is relatively little error in the scores; low reliability indicates a great amount of error. Results are usually reported in terms of a reliability coefficient. Sattler (2008) stated that tests must reach a reliability coefficient of at least .80 to be considered minimally reliable. Coefficients of .90 or above are considered the most desirable (Salvia, Ysseldyke, & Bolt, 2010; Sattler, 2008).

Hammill and Larsen (1996) examined three different sources of error variance to report reliability for the TOWL-3. Four different types of reliability coefficients were reported from analysis of content sampling, time sampling, and interscorer differences. Content sampling refers to the internal consistency reliability of test items. This was done by comparing items within the same test to determine which items correlate with each other. A derivation of Kuder-Richardson formula was used to determine the coefficient

alpha. The coefficient alphas for the Spontaneous Writing composite for 7 year olds was *r* = .89 for Form A and *r* = .90 for Form B. Another procedure was used to estimate error due to content sampling. Alternate form reliability is determined when both forms of a test are given during one testing sessions and the correlation between two forms is a reliability index. Alternate-form reliabilities for 7-year olds was *r* = .83 for the composite of the Spontaneous Writing subtest (Hammill & Larson, 1996). The individual subtests within Spontaneous Writing ranged from *r* = .60 to .87. The time sampling reliability was also computed for second grade. Time sampling examines the extent to which a student's test performance is constant over time. A test-retest correlation method was used to test the TOWL-3's time sampling error. The composite for Spontaneous Writing was .83 and .87 on the two forms of the test. Individual subtests ranged from *r* =. 76 to. 84. Reliability between scorers was the last form of reliability examined. To determine interscorer reliability, two members of the TOWL-3 publisher's staff independently scored 38 TOWL-3 protocols drawn at random from the normative sample. Reliability coefficients were not reported by individual age or grade levels, but as overall coefficients. For the composite Spontaneous Writing, *r* = .92 was reported as the mean. The individual subtests ranged for *r* = .83 to .92.

Validity. Validity refers to the degree to which different types of accumulated evidence support the intended interpretation and use of the test scores (American Educational Research Association, 2014). The authors of the TOWL-3 explored the content-related, criterion-related, and construct-related validity evidence (Hammill & Larsen, 1996). Content-related evidence examines the test content to determine whether it appropriately samples the behavior from the domain of the construct it is intended to

measure (American Educational Research Association, 2014). Classical item analysis was used to screen items during test development. Item discriminating power and item difficulty were examined. The discriminating power was determined by correlating each item with the total score of the subtest. The story construction validity was reported at .69 and .66 for the two forms of the test. Content-related validity was also assessed with Differential Item Functioning (DIF) analysis in order to detect biased items. Hammill and Larsen (1996) used the Delta Scores approach. The larger the correlation coefficient between Delta Scores across the groups, the smaller the bias in the test. DIF was conducted to make item comparisons between White and non-White students, male and female students, and Hispanic and non-Hispanic students. Results of the analysis were very high, with all of the coefficients falling at or above .95. This indicates little to no item bias in these areas.

Criterion-related evidence attempts to demonstrate that test scores are related to some other criterion variable that is thought to measure a similar construct (American Educational Research Association, 2014). Thus, if the TOWL-3 is a valid measure of writing, it should correlate well with other measures that are known or presumed to measure writing. In one study, Hammill and Larsen (1996) correlated 76 elementary students' TOWL-3 scores with the *Comprehensive Scales of Student Abilities* (CSSA; Hammill & Hresko, 1994). The CSSA is a teacher rating scale that measures a wide variety of school-related behaviors that includes writing, along with verbal thinking, speech, reading, handwriting, and mathematics. Moderate correlations were reported between the composite of Spontaneous Writing and the Writing Scale of the CSSA, with a correlation coefficient of .50.

Construct-related validity refers to the extent to which a test measures a theoretical construct or model. Using a process suggested by Gronlund and Linn (1990), Hammill and Larsen (1996) generated hypotheses regarding the underlying construct of the TOWL-3. The hypotheses are then verified by logical or empirical methods. Their first hypothesis was that the TOWL-3 scores would increase with chronological age, up to age 11 or 12 and then level off. This would demonstrate that the TOWL-3 scores were measuring students' improvement in writing skills as they continued to receive formal instruction in writing in their elementary years, and level off after their explicit instruction in writing ended. The authors examined the means and standard deviations for the eight subtests for the normative sample, and reported an increase in means between the ages of 7 and 12, and then the means leveled off after age 13. Correlation coefficients showing the relation of age to performance on the TOWL-3 subtests were also examined. The correlation coefficients were substantially stronger for students between the ages of 7 and 12 than for students between the ages of 13 and 17, supporting Hammill and Larsen's hypothesis that an increase in writing abilities will level off after students discontinue their formal writing instruction, which usually occurs at age 11 or 12.

The authors' second hypothesis was that the subtest scores of the TOWL-3 would correlate to a significant and practical degree, since they are all measuring some aspect of writing (Hammill & Larsen, 1996). All the raw scores from the normative sample were correlated, adjusting for age. All correlation coefficients were statistically significant at or beyond the .01 level. For Forms A and B the correlation coefficients ranged from .36 to .74 (median = .56) and from .33 to .75 (median = .56), respectively. These findings indicate strong relationships between the various subtests and support the construct

validity of the test.

The third hypothesis was that the TOWL-3 scores would differentiate between groups of students known to have average skills in writing and those students known to have poorer skills in writing. The means for two subgroups, students with learning disabilities and students with speech impairments, were compared to the means for the normative sample. The mean standard scores for the individual subtests for students with learning disabilities and students with speech impairments ranged from 7 to 8, whereas the mean standard score for the normative sample was 10. The mean composite quotient for students in the two subgroups ranged from 82 to 85, whereas the mean quotient for the normative sample was 100. This indicates that the TOWL-3 was able to successfully differentiate students with writing difficulties from the normative sample.

Hammill and Larsen's (1996) fourth hypothesis regarding the underlying construct of the TOWL-3 was that students who do well in writing, as measured by the TOWL-3, would do well in other academic subjects such as reading and math since they are all part of basic school skills. To test this hypothesis, scores from the TOWL-3 were correlated with three subscales of the *CSSA* (Hammill & Hresko, 1994) in a sample of 76 students. Correlation coefficients between the composite quotients of the TOWL-3 and the three subscales of the CSSA (reading, math, and general facts) ranged from .52 to .70 (median = .60), indicating a moderate relationship. The fifth hypothesis was that the scores from the TOWL-3 would significantly correlate with IQ scores, since writing is considered an intellectual ability. To test this hypothesis, 52 high-school students' TOWL-3 scores were correlated with scores from the Comprehensive Test of Nonverbal Intelligence (Hammill, Pearson, & Wiederholt, 1996). Resulting correlation coefficients

were all significant at or beyond the .05 level and ranged from .30 to .60 (median = .50), demonstrating a moderate relationship between TOWL-3 scores and IQ scores.

Construct validity relates to the degree the underlying traits of a test can be identified. Since all of the TOWL-3 subtests measure some aspect of writing, the authors hypothesized that the subtest scores would load on one factor (Hammill & Larsen, 1996). This one factor would be a measure of general writing ability, which would best be represented by the Overall Writing Quotient. A principal components analysis was completed on the data from the normative sample. In addition, principal components analyses were performed for specific subgroups, including males, females, Anglo-Europeans, African Americans, Hispanics, students with learning disabilities, and students with speech impairments. For every analysis that was computed, only a single factor emerged. Factor loadings were only available for the entire normative sample, in which they ranged from .40 to .80. The last hypothesis regarding the construct-related validity of the TOWL-3 was that the items of the individual subtests would correlate highly with the total subtest score. This is also known as an item's discriminating power. Eighty three percent of the resulting correlation coefficients fell in the acceptable range of .30 or higher. The authors concluded that it is highly unlikely for a test with poor construct-related validity to be composed of items that have such high discriminating powers.

Procedure

This study analyzed data collected through informal and formal assessments administered by classroom teachers. The district writing assessment and the CBM-W metrics were used as independent variables compared to the criterion measure, TOWL-3. Therefore, the participants in this study completed three different writing samples. The

district writing assessment was administered during the last week of January. The teachers received administration training on February 2nd on both the TOWL-3 and CBM measures. The TOWL-3 was administered during the week of February 9th, and the CBM measure was administered the following week. The teachers used a coding system to identify each student that maintains confidentiality from this researcher (ex: 4C). The passages from the TOWL-3 and CBM measures were coded the same way before scoring to maintain student confidentiality. The students' age and gender was also be identified and coded by the classroom teacher. I scored all of the TOWL-3 and CBM passages. In addition, 15% of the samples were scored by a practicing teacher and current graduate student to determine interrater reliability. This additional scorer participated in a training session and completed practice problems until satisfactory performance is achieved on both the TOWL-3 and the CBM measures. After the measures were scored and analyzed for this validity project, the extant data were used for the study.

The district writing assessment is administered three times a year in September, January, and May. This study analyzed the January data. Before teachers began scoring their students' writing, they reviewed exemplar samples for each level of the rubric to calibrate their evaluation process. Teachers scored their own classroom's samples, and discussed any scoring concerns they had with fellow teachers. Each student was given a timed CBM measure. Students were instructed to think for one minute and then write for 3 minutes to the prompt, "If I could fly I would go…" Students were encouraged to write for the entire time that was given.

Data Analysis

Multiple regression analysis was used in this study to measure the relations between CBM-W indices and the TOWL-3. After these relations are determined,

information from the district writing assessment was added to the regression procedure to determine if these scores increase this relation. Last, based on the findings from the first two analyses, data were analyzed regarding the relation of scores using both measures (CBM and analytic quality rubric) on the same writing sample with the TOWL-3. Before describing these procedures, a review of the conditions surrounding the use of multiple regression is necessary.

The first assumption is that responses must be multivariate normally distributed. Before this can be analyzed, it is necessary to analyze the data set through procedures to determine univariate and bivariate normality. Univariate and bivariate procedures are necessary, but insufficient indications of multivariate normality (Burdenski, 2000). In univariate procedures, the normal curve is determined by a mathematical equation that uses the mean and standard deviation values to generate the statistics known as skewness and kurtosis. Skewness refers to the degree of symmetry of the distribution. Kurtosis refers to the shape of the distribution against the normal distribution. It is determined by comparing relative height to width. Skewness should be within the range -2 to +2 and the kurtosis values should be within -7 to +7 (West, Finch, & Curran, 1995). Histograms were generated for each dependent variable and compared to the normal curve. The Shapiro-Wilk's W test was used as the statistical test for normality for each dependent variable. According to Razali and Wah (2011) the Shapiro-Wilk's has the best power for a given significance. Scatterplots of the variable pairs were generated to examine bivariate normality (Burdenski, 2000). If dependent variables did not meet the requirements from these normality tests, a Tukey transformation ladder was used to re-express the variables using a power transformation (Tukey, 1977). This will produce a

linear relation that can then be used to analyze the relations between variables.

The next area to analyze is sample size. The ratio of sample size to predictor variables has to be substantial or the solution will be meaningless. The number of participants needed depends on the desired power, alpha level, number of predictors, and expected effect sizes (Tabachnick & Fidell, 2001). Green (1991) provided a simple rule of thumb for testing multiple correlations and individual predictors in regression equations that assumes a medium-size relation between the predictor and criterion variables. The suggested formula is $N \geq 50 + 8m$ (where m is the number of predictors) for testing multiple correlations. Using this rule, use of 5 predictor and criterion variables indicates that 90 participants are necessary for this study.

In addition to the simple rule of thumb, a power analysis is recommended to determine the number of participants needed in a given study (Cohen, Cohen, West, & Aiken, 2003). The power of a test represents the probability of failing to reject the null hypothesis when it is false (i.e., type II error). A power analysis helps determine if the sample size is large enough to detect a significant effect (Cohen et al., 2003; Tabachnick & Fidell, 2001). If a researcher knows the number of predictor variables, desired level of power, the significance criterion (i.e., Type I error rate), and the effect size, the sample size which is necessary to meet these specifications can be determined. Cohen (1988) has designated the R^2 values of .02, .15, and .35 for small, medium, and large effect sizes in regression analyses. He also suggested that a power value of .80 is reasonable to use when there is no other basis for setting a higher or lower power. For this study, an a priori sample size multiple regression calculator (http://www.danielsoper.com/statcalc3/calc.aspx?id=1) was used to determine if 110

participants was enough to achieve power of .80 to detect an effect size of .15 at the .05 level of significance. Results of the power analysis revealed that a sample size of 91 would be appropriate for the present study; thus, a sample size of 110 exceeds the above specifications.

Multiple regression also assumes that the predictor variables individually contribute to the prediction of the criterion variable. However, if one of the predictor variables is highly correlated with another predictor variable, then those variables will contribute less unique information to the prediction of the criterion variable. This is known as multicollinearity (Cohen et al., 2003). When multicollinearity is present, the regression coefficient for the highly correlated predictor will be unreliable and have a large standard error since there is little unique information from which to estimate its value. Thus, the resulting regression coefficient would be difficult to interpret (Cohen et al., 2003; Tabachnick & Fidell, 2001). One way to screen for multicollinearity is to examine the squared correlations between each of the pairs of predictor variables. If the squared correlations are close to one, potential problems associated with multicollinearity can occur. Tabachnick and Fidell (2001) suggested that statistical problems occur at squared correlations at or above .90. If multicollinearity is present, Tabachnick and Fidell (2001) and Cohen et al. (2003) suggested omitting the predictor variable that is highly correlated with the others. For this study, the squared correlations were examined to verify the absence of multicollinearity. If multicollinearity was detected through squared correlations, the predictor variable with the highest collinearity was omitted.

The residuals from regression procedures must also meet certain assumptions. Residuals are the differences between obtained and predicted criterion scores. First,

residuals are normally distributed around the predicted criterion scores (i.e., normality). To test for normality of residuals, Cohen et al. (2003) suggested researchers plot a histogram of the residuals and then overlay a normal curve with the same mean and standard deviation on that histogram. If the histogram and normal curve are similar, then the distribution of the residuals is normal. In addition, a normal probability plot can be generated to determine if the distribution of the residuals is normal. Second, the residuals should have a straight-line relation with predicted criterion scores (i.e., linearity). Graphical methods are also recommended to test for linearity. The residuals can be plotted against each predictor variable and the predicted values. Graphs are then examined for any deviation from linearity. Third, the variance of the residuals is approximately equal for all predicted values of the criterion variable (i.e., homoscedasticity; Cohen, et al., 2003; Tabachnick & Fidell, 2001). The same graphs used to detect linearity can also be used to detect homoscedasticity (Cohen, et al., 2003). When the residuals are homoscedastic, the band enclosing the residuals will be approximately equal in width at all values of the predicted criterion score. Although statistical tests may be used to test for normality, linearity, and homoscedasticity, Cohen et al. (2003) recommended using graphical methods to help identify problems. In this study, a histogram of residuals with a normal curve overlay and a set of scatterplots of the residuals against the predictor variables and the predicted criterion scores was examined to detect if there are any violations of normality, linearity, and homoscedasticity. The last assumption to be examined is the independence of residuals. Serial dependency occurs when the data are repeatedly collected from a single individual or the same sample of individuals over time (Cohen et al., 2003; Tabachnick & Fidell, 2001). The statistical

measure, autocorrelation, was assessed by the Durbin-Watson test. The value of the Durbin-Watson coefficient ranges from 0 (positive autocorrelation) to 4 (negative autocorrelation), with 2 indicating no autocorrelation. If violations were detected in these conditions, variable transformations were used.

After all of these assumptions were met or adjusted, multiple regression analyses were used to determine the relation between CBM indices and written expression as measured by the TOWL-3. Simply stated, multiple regression combines two or more predictor variables to predict a value on a criterion variable (Tabachnick & Fidell, 2001). The goal of multiple regression is to arrive at a set of regression coefficients for the predictor variables that bring their predicted values from the equation as close as possible to the values actually obtained. There are three types of multiple regression techniques: standard, sequential, and step-wise. The procedures differ in the selection order of the predictor variables and how they enter the regression equation. In standard multiple regression, all of the predictor variables enter the regression equation simultaneously (Tabachnick & Fidell, 2001). Each predictor variable is evaluated based on the unique information that it provides to the prediction of the criterion variable. In standard multiple regression, the full correlation and the unique contribution of the predictor variables are considered in the interpretation of results. Standard multiple regression assesses the relations among predictor and criterion variables and answers two fundamental questions: (a) What is the size of the overall relation between the criterion variable and the set of predictor variables? and (b) How much is each predictor variable contributing uniquely to the prediction of the criterion variable? Based on these characteristics, this study will use standard multiple regression to analyze the unique contributions of the predictor

variables, CBM indices, to the prediction of the criterion score on the TOWL-3. Sequential regression answers the following question: Does a certain predictor variable significantly add to the prediction of a criterion variable after variance due to other predictor variables is accounted for? Since this question does not align with the purpose of this study, sequential regression was not used in the initial data analysis. Cohen et al. (2003) recommended the following three conditions be satisfied when using stepwise regression: (a) the primary research goal is predictive, not explanatory, (b) the sample size is very large, and (c) the results are cross-validated with a new sample. Stepwise regression was not be used in this study because sample sizes are not large enough to meet these assumptions for the stepwise regression technique.

In standard multiple regression analysis, the association between the predictor variables and the criterion variable is measured with the multiple correlation coefficient (R^2). This is the proportion of variance shared between the criterion variable and the predictor variables. The overall inferential test (the *F* test) will measure the significance of *R*, which is the same as testing the significance of R^2. The *F* ratio is the mean square regression over the mean square residual. If the *F* test is significant, then the null hypothesis that all correlations and regression coefficients between the predictor and criterion variables are zero is rejected. *t*-tests were used to examine the significance of individual standardized partial regression coefficients (β). The *t*-test is only sensitive to the unique variance a predictor variable adds to R^2. Therefore, if two predictor variables are highly correlated the unique contribution of each will be small and may result in a nonsignficant β (Tabachnick & Fidell, 2001). In this study, the significance of β was examined. However, since β only examines the unique contribution of each predictor, the

correlations between each individual predictor variable and criterion variable were also examined. The TOWL-3 composite score for Spontaneous Writing was used for the criterion variable. Each subtest was also evaluated independently to determine if differences exist.

CHAPTER FOUR: RESULTS

Curriculum-Based Measures in writing (CBM-W), first developed by Deno and colleagues in the early 1980s, assesses a variety of fluency-based components of writing. While support exists for the use of CBM measures in the area of writing, there is a need to conduct further validation studies to investigate the utility of these measures within elementary and secondary classrooms (Fuchs, 2004). Shriner and Thurlow (2012) advocated for the incorporation of teacher ratings with CBM-W measures to increase the measure's validity. This study explored this suggestion by analyzing results from three different writing measures using statistical analysis to determine if this was viable alternative to the traditional CBM-W measures. The following research questions were addressed in this study:

1. Which CBM metrics (TWW, words spelled correctly [WSC], CWS, percentage of words spelled correctly [%WSC] and percentage of correct word sequences [%CWS]), alone or in combination, best predict writing performance as measured by the TOWL-3?
2. What is the relation between the scores on the district writing assessment and the TOWL-3?
3. What impact does adding an analytic rubric score generated from a classroom writing activity have on the CBM metrics' prediction of TOWL-3 performance.

Interrater Reliability

Because writing scores are based on subjective scores by the reviewing, it is important to compute interrater reliability (Pierangelo & Giuliani, 2012). As part of this study, interrater reliability was computed for two of the writing measures used in this

study: CBM-W and TOWL-3. The CBM-W metrics of total words written (TWW), words spelled correctly (WSC), correct word sequences (CWS), percentage of words spelled correctly (%WSC) and percentage of correct word sequences (%CWS) were scored from a 3-minute writing sample. When using these indices, the scorer is examining the production and accuracy of writing. The TOWL-3, a norm-referenced assessment, has three different subtests that are used to derive an Overall Writing Quotient. The three subtests are Contextual Conventions, Contextual Language, and Story Construction. While using this assessment, the scorer is analyzing both the conventions of writing and the content of the story.

Table 3

Interrater Reliability for CBM-W Indices and TOWL-3 Spontaneous Writing Quotient

Measures	Coefficients
Total Words Written	.996
Words Spelled Correctly	.994
Percentage of Words Spelled Correctly	.980
Number of Correct Word Sequences	.995
Percentage of Correct Word Sequences	.983
TOWL-3 Spontaneous Writing Quotient	.982

Note. Coefficients derived from 17 randomly chosen samples from the 103 cases.

The participants in this study were 110 second grade students in a public school district in Eastern Iowa. Of the 110 participants, 103 of them completed all three assessments that were included in this study. The seven students that did not complete all three assessments were not included in the analysis. A random number generator was

used to identify 17 writing samples from both the CBM-W and TOWL-3 to determine interrater reliability. Each sample was initially scored by the primary researcher and then scored by a trained practicing teacher and current graduate student with experience administering and scoring CBM-W measures. Each CBM-W metric and the Overall Spontaneous Quotient scores were compared through correlation analysis. Average interrater reliability between the primary researcher and independent rater was high. All correlations were above .99 for the CBM countable indices and .98 for the percentage indices. The correlation was also .98 for the Spontaneous Writing Quotient of the TOWL-3 (see Table 3). Salkind (2006) recommends coefficients of .90 or higher for interrater reliability. When analyzing interrater reliability of writing samples, an exact percentage of agreement of 70 percent or better indicates reliable scoring (Brown et al., 2004).

Conditions

Prior to analyses, the CBM indices, the District Writing Assessments, and the TOWL-3 Overall Writing Quotient were examined through various scatterplots and statistical equations for accuracy of data entry, absence of outliers, absence of multicollinearity, and fit between their distributions and the assumptions of multivariate analysis. Using Cohen and colleagues' (2003) recommendations, none of the scores had a discrepancy value equal to or greater than four and the Cook's D statistics were all below one. Therefore, no extreme outliers were detected or deleted from the analysis. The mean and standard deviation values of each of the variables were used to generate skewness and kurtosis values (see Table 4). These statistics were all within the specified levels of -2 to +2 for skewness and -7 to +7 for kurtosis (West et al., 1995).

Table 4

Descriptive Statistics on CBM Indices, TOWL-3, and District Assessment

Measure	Mean	SD	Kurtosis	Skew
TWW	26.03	11.02	.38	.07
WSC	22.56	10.79	.49	-.00
CWS	19.24	11.31	.68	-.15
%WSC	.85	.11	-.95	.77
%CWS	.66	.21	-.21	-.97
TOWL- CC	9.86	2.07	.84	1.44
TOWL- CL	10.49	1.82	.54	1.11
TOWL- SC	11.01	1.46	.58	.83
TOWL- SWQ	103.01	9.85	.89	1.09
DISTRICT	6.69	2.01	-.72	-.48

Note. TWW = total words written. WSC = number of words spelled correctly. %WSC = percentage of words spelled correctly. CWS = number of correct word sequences. %CWS = percent of correct words sequences. TOWL-CC = Test of Written Language- Contextual Conventions. TOWL-CL = Test of Written Language- Contextual Language. TOWL-SC = Test of Written Language- Story Construction. TOWL-SWQ = Test of Written Language- Spontaneous Writing Quotient. DISTRICT = District Writing Assessment.

The data were then screened for multicollinearity. Correlations between TWW, WSC, CWS, %WSC, and %CWS were completed and squared correlations were analyzed (see Table 5). Even though all of the variables were statistically significant

when correlated, only WSC and TWW (R^2 = .94) had a correlation over .90. WSC also had a strong correlation with CWS (R^2 = .85). To reduce estimation problems as a result of the redundancy among predictor variables, WSC was omitted from the study (Cohen et al., 2003; Morrow-Howell, 1994). Once the WSC variable was removed from the analysis, all squared correlations were below .90.

Table 5

Intercorrelations between CBM-W and TOWL-3 Spontaneous Writing Quotient

	TWW	WSC	CWS	%WSC	%CWS	TOWL- SWQ
TWW	1.00	.970**	.833**	.313**	.214*	.317**
WSC		1.00	.921**	.507**	.397**	.366**
CWS			1.00	.624**	.653**	.466**
%WSC				1.00	.853**	.308**
%CWS					1.00	.349**
TOWL-SWQ						1.00

Note. **Correlation is significant at the 0.01 level. *Correlation is significant at the 0.05 level. TWW = total words written. WSC = number of words spelled correctly. %WSC = percentage of words spelled correctly. CWS = number of correct word sequences. %CWS = percent of correct words sequences. TOWL-CC = Test of Written Language- Contextual Conventions. TOWL-CL = Test of Written Language- Contextual Language. TOWL-SC = Test of Written Language- Story Construction. TOWL-SS = Test of Written Language- Standard Score.

Figure 1

Histogram of Standardized Residuals

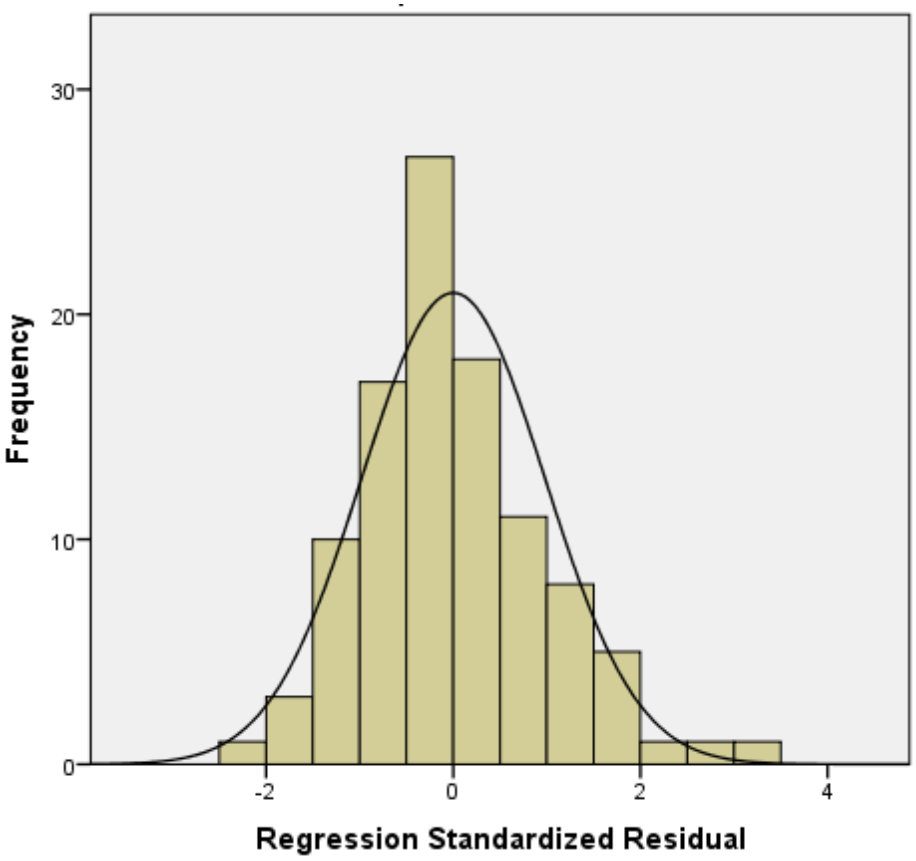
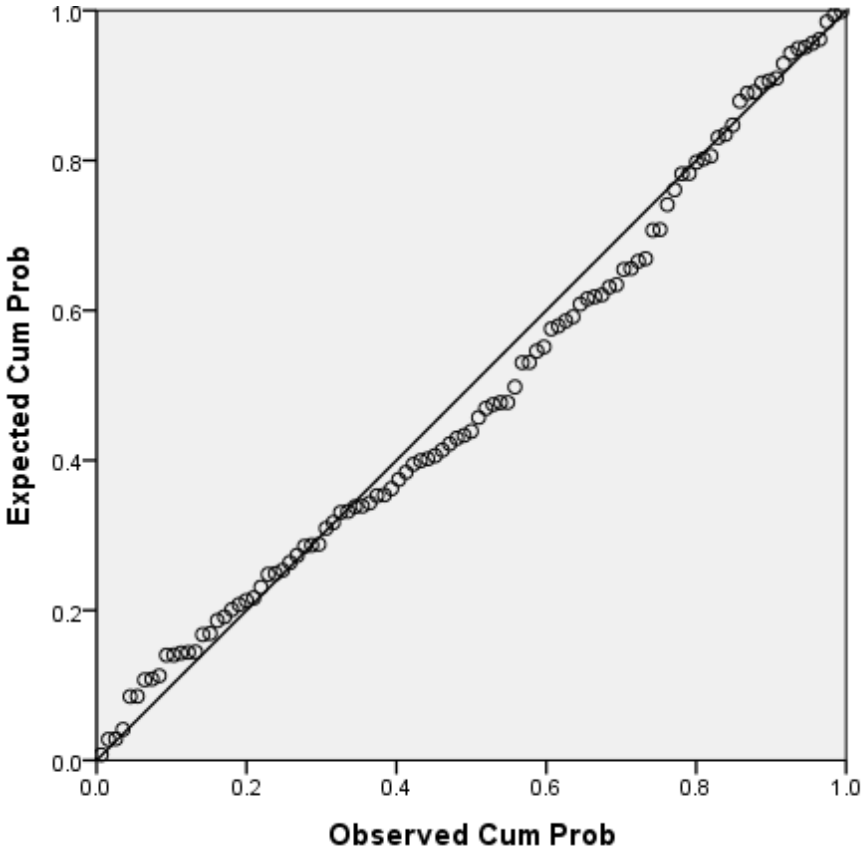


Figure 2

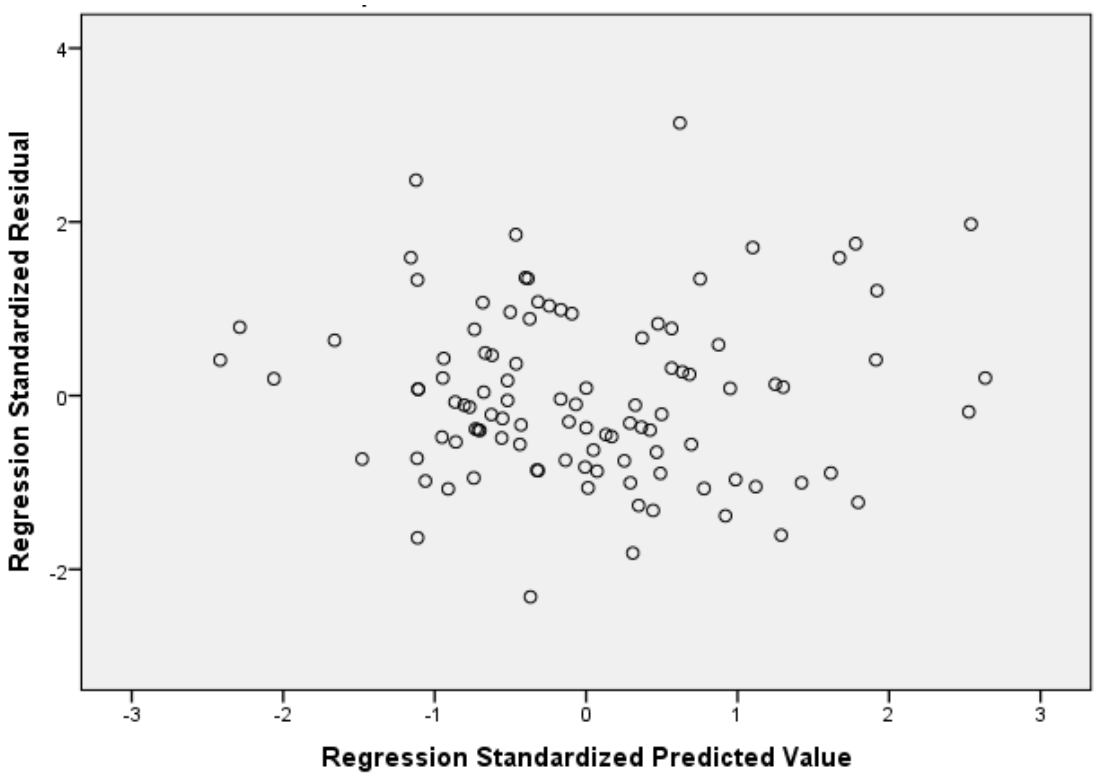
Normal P-P Plot of Regression Standardized Residuals



Last, the residuals were examined to determine if they met the assumptions of normality, linearity, homoscedasticity, and independence. In regards to the assumption of normality, the distribution of residuals was relatively normal (see Figure 1). A normal probability plot revealed a nearly straight line (see Figure 2). This indicates the assumption of normally distributed residual error was met. A set of scatterplots examining the linear relations between the predictor and criterion variables and the residuals and the predictor variables were inspected (see Figure 3).

Figure 3

Scatterplot of Standardized Predicted Values



There was no substantial departure from the assumption of linearity. However, when examining the individual scatterplots of the residuals and the predictor variables for homscedasticity, the graphs revealed that both percent measures (i.e., %WSC and %CWS) did not evenly spread out. However, percent measures often have distributions that are rectangular instead of normal (Cohen et al., 2003; Tabachnick & Fidell, 2001). Therefore, an arcsine transformation was used to flatten the distribution and stretch out both tails. This is a procedure to normalize proportional distributions (Howell, 2002). When this transformation was applied to %WSC and %CWS, the overall results were not significantly different. Given the minor differences in results, the untransformed variables were used for ease of interpretation. The last assumption that was examined was

independence of residuals. Serial dependency occurs when the data are repeatedly collected from a single individual or the same sample of individuals over time (Cohen et al., 2003; Tabachnick & Fidell, 2001). The statistical measure, autocorrelation, can be assessed by the Durbin-Watson test. The value of the Durbin-Watson coefficient ranges from 0 (positive autocorrelation) to 4 (negative autocorrelation), with 2 indicating no autocorrelation. In this study, this Durbin –Watson coefficient was 2.063, indicating no autocorrelation, or independence of residuals.

Results by Research Questions

Descriptive statistics are presented in Table 5. The mean, standard deviation, kurtosis, and skewness are presented for the CBM-W metric (TWW, WSC, CWS, %WSC, %CWS), the TOWL-3 subtests (Contextual Conventions, Contextual Language, Story Construction), and the TOWL-3 composite score for Spontaneous Writing Quotient. The mean for the Spontaneous Writing Quotient was 103.23, which is slightly above the 100 standard score from the TOWL-3. Local normative scores indicate that a TWW mean of 26.50 is just above the 54th percentile while a CWS mean of 19.58 is just above the 65th percentile. The large standard deviations indicate strong variability in these measures, but neither significant skew nor kurtosis was evident. Each research question was answered using the results from statistical analysis of the data.

Research question 1. Which CBM metrics (TWW, words spelled correctly [WSC], CWS, percentage of words spelled correctly [%WSC] and percentage of correct word sequences [%CWS]), alone or in combination, best predict writing performance as measured by the TOWL-3? Pearson correlations were used to compute intercorrelations among the CBM indices and the TOWL-3 Spontaneous Writing Quotient (see Table 3). This indicated significant correlations between each of

the CBM-W measures and the TOWL-3 Spontaneous Writing Quotient ($p < .001$). When compared with the other CBM-W measures, CWS had the highest correlation with the Spontaneous Writing Quotient ($r = .460$). %WSC was the next highest correlation at .364 and WSC at .357. %WSC and TWW were the lowest at $r = .316$ and $r = .306$.

Standard regression analysis was conducted to explore which CBM indices best predicted the TOWL-3 Overall Writing Quotient. Students' scores on TWW, %WSC, %CWS, and CWS were entered simultaneously as predictors in a regression analysis that included students' scores on the TOWL-3 Spontaneous Writing Quotient as the dependent variable. Results of the regression analysis are shown in Table 6.

Table 4

CBM-W and TOWL-3 Regression ANOVA

Model	Sum of Squares	*df*	Mean Square	*F*	Sig
Regression	2381.41	4	595.35	7.43	.000
Residual	7849.00	98	80.09		
Total	10230.41	102			

$p < .001$

The multiple correlation coefficient was significantly different from zero ($r = .482$) and the four predictor variables collectively accounted for 23% of the variance in students' TOWL-3 Spontaneous Writing Quotient, $F(4, 98) = 7.433$, $p < .001$. Only one of the predictor variables significantly contributed unique variance to the prediction of the TOWL-3 Spontaneous Writing Quotient. As seen in Table 7, CWS contributed the most unique variance ($\beta = .791$, $p < .05$).

Table 5

Summary of CBM-W Index Scores as Predictors of Overall Writing Quotient

Variable	*B*	*SE B*	*ß*	*t*	*p*
Intercept	102.056	9.195		11.099	.000
TWW	-.361	.230	-.382	-1.44	.152
CWS	.791	.315	.881	2.514	.014
%WSC	.064	14.884	.001	.004	.997
%CWS	-7.265	12.772	-.153	-.569	.571

Note. TWW = total words written. WSC = number of words spelled correctly. %WSC = percentage of words spelled correctly. CWS = number of correct word sequences. %CWS = percent of correct words sequences

Research question 2. What is the relation between the scores on the district writing assessment and the TOWL-3? To determine this relation, CBM-W measures and the TOWL-3 were used as criterion measures. Pearson correlations were computed for each CBM-W variable (see Table 8). All of the countable CBM-W indices (TWW, WSC, CWS) correlate to the district rubric at statistically significant levels ($p < .001$). With correlations ranging from .316 to .334, these correlations are of moderate strength. CWS most strongly correlated with the district rubric. %CWS was also statically significant at $p < .05$.

Table 6

Correlations between the District Writing Assessment and the CBM-W Measures

	TWW	WSC	CWS	%WSC	%CWS
DISTRICT	.316*	.317**	.334**	.165	.197*

Note. . **Correlation is significant at the 0.01 level. *Correlation is significant at the 0.05 level. TWW = total words written. WSC = number of words spelled correctly. %WSC = percentage of words spelled correctly. CWS = number of correct word sequences. %CWS = percent of correct words sequences. DISTRICT = District Writing Assessment.

Pearson correlations were computed between the District Writing Assessment and each of the TOWL-3 subtests and the Spontaneous Writing Quotient (see Table 9). The correlation with the Story Construction subtest was statistically significant ($p < .01$). With a correlation coefficient of .291, this relation is at the low-moderate level. The Spontaneous Writing Quotient and Contextual Conventions were also statistically significant ($p < .05$) with overall weak relations of .239 and .197 respectively. The correlations with Contextual Language and %WSC were positive, but not at the significant level.

Table 7

Correlations between the District Writing Assessment and the TOWL-3 Subtests

	TOWL-CC	TOWL-CL	TOWL-SC	TOWL-SWQ
DISTRICT	.192*	.151	.291**	.239*

Note. . **Correlation is significant at the 0.01 level. *Correlation is significant at the 0.05 level. TOWL-CC = Test of Written Language- Contextual Conventions. TOWL-CL = Test of Written Language- Contextual Language. TOWL-SC = Test of Written Language- Story Construction. TOWL-SWQ = Test of Written Language- Spontaneous Writing Quotient. DISTRICT = District Writing Assessment.

Research question 3. What impact does adding an analytic rubric score generated from a classroom writing activity have on the CBM metrics' prediction of TOWL-3 performance? When the District Writing Assessment was added to the CBM-W regression, the prediction increased slightly (r = .493) and contributed to 24% of the variance, F (5, 97) = 6.243, $p < .001$ (see Table 10). This was not a statistically significant change.

Table 8

CBM-W and District Writing Assessment and TOWL-3 Regression ANOVA

Model	Sum of Squares	*df*	Mean Square	*F*	Sig
Regression	2490.67	5	498.13	6.24	.000
Residual	7739.74	97	79.79		
Total	10230.41	102			

p < .001

CHAPTER FIVE: DISCUSSION

CBM was designed to be a technically adequate method for collecting information in reading, math, and writing. Although evidence for reliability and validity has been researched in all three areas, most of the CBM literature has focused on reading and math. The extensive CBM data in reading support the use of one measure, oral reading fluency, as an indicator of performance and progress in reading for elementary school students. This index has been shown to strongly correlate with a variety of criterion measures across many studies, which included different participants, methods, materials, and researchers (Wayman, Wallace, Wiley, Ticha, & Espin, 2007). There is not one CBM-W index that has been consistently shown to be technically sound or theoretically appropriate (McCaster & Espin, 2007; Saddler & Asaro-Saddler, 2013). Concerns exist regarding the use of these measures for screening, progress monitoring, and eligibility decisions (Coker & Ritchey, 2010; Espin et al., 2004; Espin et al., 2005; Jewell & Malecki, 2005; McMaster et al., 2009).

Of all of the possible CBM-W measures, CWS has been found to have strong correlation with holistic measures, but is not as sensitive to growth (Costa et al., 2012; Gansle et al., 2002; Parker et al., 1991b; Ritchey & Coker, 2013). TWW is more sensitive, but it is weak in validity, especially in recent studies (Amato & Watkins, 2011; Costa et al, 2012). Also, the measures do not contain the high predictive validity that is seen in reading and math CBM measures (McMaster & Espin, 2007). For example, the present study yielded low-moderate (r = .306) to high-moderate (r = .466) validity coefficients between CBM writing indices and the TOWL-3. In contrast, reading CBM validity coefficients have consistently been in the .70s or above. Not only do CBM-W measures lack the high validity of the reading measures, there is also a lack of face

validity from teachers (Foegen, 2001; Gansle, Gilbertson, & VanDerHeyden, 2006). A survey by Gansle and colleagues (2006b) found that many teachers believe that the CBM-W measures cannot capture all of the subtle nuances that contribute to quality writing. Therefore, teacher ratings provide higher levels of face validity.

In an attempt to produce writing measures that better align to validated writing assessments, Shriner and Thurlow (2012) suggested that a combination of measures (both quantitative and qualitative) might be needed. The purpose of this study was to combine CBM metrics with a teacher rating that used an analytic quality rubric. The assumption was that the addition of a quality component would increase the validity of measures that just focus on production and accuracy of writing mechanics. The following research questions guided this study:

1. Which CBM metrics (TWW, words spelled correctly [WSC], CWS, percentage of words spelled correctly [%WSC] and percentage of correct word sequences [%CWS]), alone or in combination, best predict writing performance as measured by the TOWL-3?
2. What is the relation between the scores on the district writing assessment and the TOWL-3?
3. What impact does adding an analytic rubric score generated from a classroom writing activity have on the CBM metrics' prediction of TOWL-3 performance?

Two writing samples from second grade students were scored to determine if adding a teacher rating score to CBM metrics would increase the concurrent validity when compared to an established validated measure. The timed writing passages were scored using the CBM-W metrics of TWW, WSC, CWS, %WSC, and %CWS. The

students also wrote passages for the district writing assessment. This assessment was scored using an analytic rubric with three components: content/ideas; organization; and word choice. The total score from these components determined the teacher rating score. The scores from the CBM-W metrics and the teacher rating score were then compared to the TOWL-3 (Hammill & Larsen, 1996) Spontaneous Writing subtests. All three of the writing assessments were administered by the classroom teachers in a public school in Eastern Iowa. I scored the CBM-W samples and the TOWL-3 Spontaneous Writing sample.

A major criticism of scoring writing samples to assess students' writing achievement has been the low levels of reliability between different raters (Barkaoui, 2011; Graham et al., 2011; Harsch & Martin, 2013). When scoring writing, raters have different perspectives and opinions as to what constitutes quality writing, which introduces additional sources of variance. The countable indices used in CBM-W measures have the often-cited advantage of reducing this variance. The results from this study also corroborate to this high level of reliability. In this study, all CBM indices were scored reliably following training. Average interrater reliability between the primary researcher and independent raters was high and above .98 for the CBM indices. In additional, the interrater reliability between the primary researcher and independent rater was .98 for the Spontaneous Writing Quotient on the TOWL-3.

The results of the present study provide only modest support for the use of CBM-W. All of the predictors provided statistically significant correlations to the TOWL-3, but the correlations were not strong. After WSC was eliminated from because of multicollinearity concerns, the four CBM indices accounted for only 23% of the variance

in TOWL-3 scores.

Research Question 1

Which CBM metrics (TWW, words spelled correctly [WSC], CWS, percentage of words spelled correctly [%WSC] and percentage of correct word sequences [%CWS]), alone or in combination, best predict writing performance as measured by the TOWL-3? All of the metrics had statistically significant correlation to the TOWL-3 Spontaneous Writing Quotient. This indicates that there is less than 1% likelihood that these correlations are explained by chance. Even though it is evident that there is a relation between these metrics, it is important to examine the magnitude of these relations. While looking to the individual metrics, CWS had the greatest correlation to the TOWL-3 Spontaneous Writing Quotient (r = .466). This was significantly higher than any other metric. These findings align with other research regarding the use of CWS at the elementary level. Videen et al. (1982) compared scores using CWS with students in third through sixth grade. The CWS scores were compared to the TOWL Raw Score and a holistic rating. The correlations were .69 and .85 respectively. In Parker's study (1991b) that pulled out individual grade level scores, CWS had a correlation of .60 when compared to a holistic rating with second grade students. The study by Jewell and Malecki (2005) also produced a strong correlation of .57 when compared with the SAT Language test.

WSC had a moderate correlation of r = .366. However, CWS and WSC had a large degree of collinearity (r = .921). This is not surprising since these metrics are measuring similar traits. When computing WSC, the scorer counts the total number of words that are spelled correctly. When using CWS the scorer also analyzes spelling by counting the number of correct sequences between words, but grammar and mechanics

are also included. This high level of collinearity indicates that the addition of these conventions maintains some similarity with accurate spelling, but that combining all three components into the CWS metric provides the highest level of predictive ability. This finding is consistent with other studies that analyzed both CWS and WSC at the elementary grades. Gansle and colleagues (2002) found that CWS outperformed WSC when comparing these scores with the ITBS Total Subscale with third graders. CWS had a stronger correlation of .43 compared to the WSC correlation of .24. Jewell and Malecki (2005) also found a significant difference between CWS and WSC with second graders. CWS had a strong relation at .57, but WSC was only .38 when compared with the SAT Language test. Even though WSC is easier to compute, the extra time necessary to score CWS adds to the predictive power.

The metric %CWS also correlated at the moderate level (r = .349). Since this metric involves looking at the percentage of CWS, the original CWS metric must be computed to obtain this percentage. Based on this and the much stronger correlation seen with CWS, it would only be logical to choose CWS instead of %CWS.

Even though all of the CBM-W metrics had statistically significant correlations with the TOWL-3 Spontaneous Writing Quotient, TWW and %WSC had the lowest levels of correlation. TWW, which is a simple count of the total words written in the sample, is consistently shown to be an adequate measure that does not have the same validity of the other CBM-W measures. Studies by Gansle et al. (2002), Gansle et al. (2004) and Jewell and Malecki (2005) all produced correlations less than .25 when TWW was compared to standardized assessments. Jewell and Malecki (2005) even found a negative correlation (-.14) for students in sixth grade. %WSC, which only takes into

account the accuracy of spelling and not the production, had the lowest correlation with the TOWL-3 Spontaneous Writing Quotient. If evaluating spelling, WSC has a higher predictive ability.

In addition to analyzing individual correlations, a multiple regression was conducted. WSC was removed from the multiple regression analysis because the metric had high multicollinearity with TWW and CWS. When the remaining four metrics were placed in the equation, the correlation only slightly increased beyond what it was for CWS. The correlation increased from $r = .466$ to $r = .482$, increasing the explained variance by only 1.3%. The only metric that provided a statistically significant unique variance was CWS. Since CWS is a composite metric that takes into account components from two other metrics, production from TWW and spelling from WSC, it is not surprising that adding these measures to the regression added very little validity. This further reinforces that CWS, when compared with the other metrics, provides the best prediction of scores on the TOWL-3 Spontaneous Writing Quotient

Research Question 2

What is the relation between the district writing assessment and validated writing measures? The district writing assessment that was used as a measure in this study focused on the compositional quality of the writing. The essays were scored using an analytic rubric that had three components: content/ideas, organization, and word choice. While assessing content/ideas, teachers checked if the essay responded to all parts of the prompt. The teachers also evaluated how the topic and ideas were communicated through the use of details or facts. While assessing organization, the teachers focused on sentence structure and an organized essay including a clear introduction, body, and conclusion with supporting details. The last component, word choice, evaluated the use of adjectives,

nouns, verbs, and words to show feelings, and sentence fluency. Even though the teachers evaluated the papers for conventions, spelling, and presentation, those components were not included in the rubric score and were not used to determine if students were proficient writers.

Since the focus was placed on overall content and organization of writing, not on spelling, grammar, usage, capitalization, or punctuation, it is not surprising that the rubric score aligned most closely with the TOWL-3 Story Construction subtest ($p < .01$). Items on this subtest include the components of plot and flow of the writing. Even though these assessments were both focused on quality of writing, the correlation was still considered weak ($r = .291$). The different genres of writing may explain some of this variation. On the TOWL-3, the students were asked to write a narrative story based on a picture prompt while the district assessment was an expository piece from a verbal prompt. The correlation with the composite Spontaneous Writing Quotient was significant, but only at the .05 level. The correlations with Contextual Conventions and Contextual Language subtests were not significant. This indicates that these assessments were measuring different writing skills. This also further validates that the district assessment does not assess conventions or language usage, which is part of the foundation of the rubric.

To further investigate the use of this measure, correlations were also evaluated with CBM-W metrics. The three countable indices (TWW, WSC, and CWS) all had statistically significant correlations ($p < .001$), ranging from .316 to .334. The correlations, however, were not as high with the percentage metrics (%WSC and %CWS) with correlations of .165 and .197 respectively. Even though this assessment does not focus on conventions, it does indicate that there are relations between these countable

indices and the district assessment. The students need to write enough content for it to be evaluated properly. These are skills that are also assessed in the CBM-W countable indices, and it would then explain why that same level of relation was not seen in the percentage metrics. This study indicates that this district assessment has some concurrent validity when compared to these measures. Even though the correlations are weak to moderate, it indicates that it is a unique assessment that can be used to assess different components of writing that are not evaluated in these other measures.

Research Question 3

What impact does adding an analytic rubric score generated from a classroom writing activity have on the CBM metrics' prediction of TOWL-3 performance? It has been hypothesized that adding a quality measure will increase the validity of CBM-W measures when comparing these scores to standardized assessments (Shriner & Thurlow, 2012). This study added a second writing measure that used an analytic rubric focused on the quality of the composition. The additional assessment score only slightly increased the prediction when added to the CBM-W measures. It also did not uniquely contribute to the TOWL-3 scores.

There are many possible hypotheses to explain this. At the second grade level, it is necessary for students to write enough information to gain an adequate score on the rubric. If a student writes very little on both measures, very similar low scores will be apparent. Therefore, the rubric is not identifying unique contributions from the CBM measures. The TOWL-3 Spontaneous Writing Quotient is derived from three subtests. Two of those subtests are focused on grammar, language usage, spelling, and capitalization. These areas are the components of CWS. The Story Construction subtest

did have a stronger correlation to the District Writing Assessment, indicating that similar skills were assessed in these two tests. These findings need to be interpreted in light of some potential limitations.

Limitations and Future Research

A limitation to the study is that all students who participated were from one grade level in a single school district with a relatively small number of participants. Thus, generalizations beyond this sample should be limited to second grade students from similar demographics. This research should be replicated with different grade levels. Even though the analytic rubric did not significantly increase the validity of CBM-W metric at second grade, it may have a significant impact as the grades increase. Students in the upper elementary grades are receiving instruction on specific elements of writing: clarity; support of main ideas and subpoints; organization and development (Miller, 1995). Mechanics, including grammar, spelling, punctuation, capitalization, paragraphing, and sentence structure are addressed, but not at the same level as early grades. At these upper grades, an analytic rubric that addresses these components of writing might provide the necessary validity. The validity of CBM-W decreases as students progress into middle school (Amato & Watkins, 2011; Espin et al, 2000; Espin et al., 2005). In addition, replication with a diverse population regarding socio-economic status and race should be considered.

Another limitation of this study was that the writing samples were obtained using different administration protocols. The CBM-W sample was timed for three minutes, and students had 15 minutes to write during the TOWL-3. On the district assessment, students were given up to 45 minutes to compose their essay. Some students might have difficulty writing under time constraints of three minutes or even fifteen, but would feel more

comfortable having an extended amount of time. The extended amount of time also allows for students to make revisions, which would impact the quality of their writing. Students were also able to use planning worksheets during the district assessment, which might also impact the quality of their writing which was not seen on the TOWL-3. In addition to the differences in administration, there were also differences in the scoring metrics. For example, the district writing assessment only had a possible range of 2-9, while the TOWL-3 transformed scores ranged from 21-46, CBM metrics ranged from 6-54. This lack of range variability in the district writing assessments could have influenced the low correlations.

There were also differences in the genre of writing in each assessment. Students were asked to write a narrative story during the TOWL-3 administration that had a beginning, middle, and end with characters. The district assessment and CBM-W protocols asked for an expository essay, but students could have addressed the CBM-W prompt creatively. That prompt asked student if they could fly, where would they go. Some students answered the question by describing trips to Hawaii or a local amusement park. Others addressed the prompt in a fantastical matter describing a visit to characters from a movie or traveling back in time. These variations in administration may have limited the relations between these different measures. Genre knowledge, including task schemas that specify how to carry out particular writing tasks, is a factor that impact students' writing achievement (Olinghouse & Graham, 2009). However, using two different types of writing samples as the independent variable was a key component of this study since the focus was to find if adding another writing sample would increase the validity of CBM-W while using the TOWL-3 as the criterion.

A possible direction in the future would be to eliminate the extra writing sample and use a quality rubric on the CBM writing sample. A suggestion from Hessler and Konrad (2008) described using the same writing sample to measure both the CBM components and quality. In this example, students would write for the determined amount of time (e.g., three minutes) and then make a mark on their papers. They would then finish their composition. The assessor would score the entire essay for quality, but only the section from the first three minutes for the CBM measures. This method may decrease the differences between different writing genres and administration procedures. Future research should also pull the different components apart from the analytic rubric to determine if a specific area may provide more validity than the entire rubric.

Last, this study only investigated the concurrent validity of these writing measures. The CBM and criterion measure were given at one point in time. To examine the suitability of progress monitoring, the CBM-W measures would need to be administered multiple times within a confined amount of time. Then, the technical features of the slope would need to be analyzed to determine if increasing CBM scores are associated with increased performance in writing achievement (Fuchs, 2004). In addition, this study did not examine proficiency levels or cut-scores. These levels are determined by analyzing the accuracy in identifying individuals who will pass the future criterion measure. If decisions, such as special education eligibility, are going to be made based on performance of CBM metrics, these proficiency levels must be identified (Johnson, Jenkins, Petscher, & Catts, 2009).

Research is still needed in the three stages outlined by Fuchs (2004) to establish the utility of CBM-W. This study focused on the first stage: examining the technical

features of the static score. It analyzed two static scores and correlated these to an outcome measure (Glover, 2010b). The focus was to determine if the validity of CBM-W measures could be increased by adding a quality component, but it was just analyzing static scores. The majority of research on CBM-W measures has been at Stage 1 because a measure has not been found that provides to the necessary validity for use in screening and progress monitoring purposes. Stage 2 involves examining technical features of the slope. Since each weekly test is comparable in difficulty and conceptualization, slopes can be used to quantify rate of learning (Fuchs, 2004). Very little research has been conducted at Stage 2. It may be necessary for researchers to recognize that we might not achieve the levels of validity in writing that we have in reading. It may be time to take the measure that shows the greatest validity, CWS, and use it to begin analyzing the slope at various grade levels. The analysis of the slope would also lead to identification of proficiency levels or cut scores that would identify those students most at-risk for writing difficulties. The last stage then examines the instructional utility of the measure. Research needs to be conducted that analyzes how teachers use CBM-W to make instructional decisions and remediate problem areas.

Implications for Practice

Implications from this study apply to both school districts and classroom teachers. Regarding school districts, it is important to collect validity data related to district assessments. Even though the district assessment used in this study had some relative validity, it was only at the weak-moderate level when compared with the TOWL-3 Spontaneous Writing Quotient. It was only when compared to the Story Construction subtest of the TOWL-3 that moderate relations were found. This should serve as a caution to districts creating their own rubrics and writing assessments without examining

the validity to established writing assessments, such as the TOWL-3.

This study also demonstrates how some students score differently on these three different writing samples. Not only does this include the different scoring approaches of quality and quantity, but also different administration procedures such as time given, type of prompt (written, picture, oral), and genre or writing style. Since the correlations were not high between the three assessments, it further demonstrates the need for schools to use multiple pieces of information when making determinations for special programs.

Implications were also found for classroom teachers. CBM-W measures provide valid information for teachers, especially the CWS metric. Not only did CWS correlate higher to the outcome measure than the other CBM-W measures, it also had a higher correlation than the district writing assessment. In this study, all five of the CBM-W measures had higher correlations than the district assessment. Even though teachers generally prefer a teacher rating instead of CBM-W (Gansle et al., 2006b), this study should provide some necessary face validity to CBM-W for classroom teachers. Teachers should be able to trust the results they receive from CBM-W measures and recognize that these measures are monitoring similarly to validated assessments, such as the TOWL-3.

This study should also demonstrate to teachers the need to provide a balanced approach to both assessing and teaching writing. This study indicates that there is a place for both writing mechanics and the quality components in writing composition. The results from this study indicate that production and writing mechanics are predictive traits in second graders' writing. Because of the higher correlation seen with CWS when compared to WSC, it also indicates that the other areas of mechanics are very important. This would include capitalization, punctuation, grammar, and usage. Even though the

quality components of ideas, organization, and word choice are important, at second grade it may be just as important that students are able to use correct mechanics and string together enough content to be accurately assessed.

Teachers may want to consider using a quality rubric with students that have either overinflated or underinflated CBM scores. When a student writes very simple but accurate sentences, the CBM-W score will be higher. A quality rubric would identify the lack of voice or composition quality in that writing. Students may also have difficulties with spelling mechanics, but have a strong writing composition. In these extreme instances, adding in a quality rubric would provide more information that CBM-W cannot provide independently.

Conclusion

The present study examined the relations between CBM scores, a district assessment and a standardized test of writing, the TOWL-3. Results of this study were consistent with past research that shows simple fluency measures, such as TWW and WSC are sufficient measures at the lower elementary level, but the more complex CWS is a better predictor of TOWL-3 scores. These results are consistent with previous research on CWS as a general indicator of writing ability (Jewell & Malecki, 2005; Marston, 1989; Parker, 1991b; Videen et al., 1982). However, it is important to note that the percentage of CWS did not perform at the same level. Although future research may suggest a CBM measure that yields reliable and valid scores and that can be used to monitor progress, this study suggests that adding a quality component to CBM-W did not result in the level of validity needed to use this tool for progress monitoring.

REFERENCES

Alber-Morgan, S., Hessler, T., & Konrad, M. (2007). Teaching writing for keeps. *Education & Treatment of Children, 30*, 107-128.

Amato, J. M., & Watkins, M. W. (2011). The predictive validity of CBM writing indices for eighth-grade students. *The Journal of Special Education, 44*, 195-204. doi:10.1177/0022466909333516

American College Testing (2014). *The Condition of College & Career Readiness 2014.* Retrieved from http://www.act.org/research/policymakers/cccr14/.

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

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice, 17*, 199-214.

Applebee, A. N., & Langer, J. A. (2011). A snapshot of writing instruction in middle schools and high schools. *English Journal, 100*, 14-27.

Atwell, N. (2002). *Lessons that change writers*. Portsmouth, NH: Firsthand/Heinemann.

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74*, 29-58.

Barenbaum, E. (1987). Children's ability to write stories as a function of variation in task, age, and developmental level. *Learning Disability Quarterly, 10*, 175-88.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, 18, 279-293. doi: 10.1080/0969594X.2010.526585

Beck, S., & Jeffery, J. (2007). Genres of high-stakes writing assessments and the construct of writing competence. Assessing Writing, 12, 60–79.

Behizadeh, N., & Engelhard, G., Jr. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing, 16*, 189-211.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: L. Erlbaum Associates.

Beringer, V., Mizokawa, D., & Bragg, R. (1991). Theory-based diagnosis and remediation of writing disabilities. *Journal of School Psychology, 29,* 57-97.

Bizzell, P. (1982). Cognition, convention, and certainty: What we need to know about writing. *Pre/text, 3*, 213-244.

Bradley-Johnson, S., & Lesiak, J. (1989). *Problems in written expression: Assessment and remediation*. New York, NY: Guilford Press.

Bromley, K. (2011). Best practices in teaching writing. In L. M. Morrow & L. B. Gambrell (Eds.), Best practices in literacy instruction (4th ed., pp. 295-318). New York, NY: Guilford

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, G., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9,* 105–121.

Burdenski, T. K., Jr. (2000). Evaluating univariate, bivariate, and multivariate normality using graphical procedures. *Multiple Linear Regression Viewpoints, 26,* 15-28.

Burgin, J., & Hughes, G. D. (2009). Credibly assessing reading and writing abilities for both elementary student and program assessment. *Assessing Writing, 14*, 25-37.

Calfee, R. C., & Miller, R. G. (2007). Best practices in writing assessment. In S. Graham; C. A. MacArthur & J. Fitzgerald (Eds.), *Best practices in writing instruction* (pp. 265- 286). New York, NY: Guilford Press.

Cherry, R. D., & Meyer, P. R. (2009). Reliability issues in holistic assessment. In B. A. Huot & P. O'Neill (Eds.), *Assessing writing: A critical sourcebook* (pp. 29-56). Boston, MA: National Council of Teachers of English.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York, NY: Harper & Row.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1966). *Cartesian linguistics: A chapter in the history of rationalist thought.* New York, NY: Harper & Row.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, L. G., & Spenciner, L. J. (2011). *Assessment of children and youth with special needs* (4th ed.). Boston, MA: Pearson.

Coker, D.L., Jr., & Ritchey, K. D. (2010). Curriculum -based measurement of writing in kindergarten and first grade: An investigation of production and qualitative scores. *Exceptional Children, 76*, 175-193.

Cole, J. C., Haley, K. A., & Muenz, T. A. (1997). Written expression reviewed. *Research in the Schools, 4*, 17-34.

Conley, M. W. (2005). *Connecting standards and assessment through literacy*. Boston, MA: Allyn and Bacon.

Costa, L. C., Hooper, S. R., McBee, M., Anderson, K. L., & Yerby, D. C. (2012). The use of curriculum-based measures in young at-risk writers: Measuring change over time and potential moderators of change. *Exceptionality, 20*, 199-217.

De La Paz, S. (2007). Best practices in teaching writing to students with special needs. *Best practices in writing instruction.* (pp. 308-328). New York, NY: Guilford Press.

Deno, S. L., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children, 48*, 368-371.

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

Diederich, P. B. (1974). *Measuring growth in English.* Urbana, IL: National Council of Teachers of English.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability*. Princeton: Educational Testing Service.

Eaves, R., & Woods-Groves, S. (2007). Criterion validity. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics.* (pp. 201-203). Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781412952644.n113

Espin, C. A., Weissenburger, J. W., & Benson, B. J. (2004). Assessing the writing performance of students in special education. *Exceptionality, 12*, 55-66. doi:10.1207/s15327035ex1201_5

Espin, C. A., De La Paz, S., Scierka, B. J., & Roelofs, L. (2005). The relationship between curriculum-based measures in written expression and quality and completeness of expository writing for middle school students. *The Journal of Special Education, 38,* 208-217. doi:10.1177/00224669050380040201

Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *The Journal of Special Education, 34,* 140-153. doi:10.1177/002246690003400303

Faigley, L. (1985). Non-academic writing: the social perspective. In L. Odell & D. Goswami (Eds.), *Writing in non-academic settings* (pp. 231–248). New York, NY: Guildford Press.

Fewster, S.. & MacMillan, P. D. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education, 23,* 149-156. doi:10.1177/07419325020230030301

Flower, L. & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication, 31,* 21-32.

Foegen, A., Espin, C., Allinder, R., & Markell, M. (2001). Translating research into practice: Preservice teachers' beliefs about curriculum-based measurement. *The Journal of Special Education, 34,* 226-236.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33,* 188-192.

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review, 31,* 477-497.

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Slider, N. J., Hoffpauir, L. D., Whitmarsh, E. L., & Naquin, G. M. (2004). An examination of the criterion validity and sensitivity to brief intervention of alternate curriculum-based measures of writing skill. *Psychology in the Schools, 41,* 291-300. doi:10.1002/pits.10166

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review, 35,* 435-450.

Gansle, K. A., Gilbertson, D., & VanDerHeyden, A. M. (2006). Elementary School Teachers' Perceptions of Curriculum-Based Measures of Written Expression. *Practical Assessment, Research, and Evaluation, 11,* 1-17.

Gearhart, M. (1992). *Writing portfolios at the elementary level: A study of methods for writing assessment.* Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.

Glover, T. A. (2010a). Key RTI service delivery components. In T. A. Glover, & S. Vaughn (Eds.), *The promise of response to intervention: Evaluating current science and practice* (pp. 7-22). New York: Guilford Press.

Glover, T. A. (2010b). Supporting all students: The promise of response to intervention. In T. A. Glover, & S. Vaughn (Eds.), *The promise of response to intervention: Evaluating current science and practice* (pp. 1-6). New York: Guilford Press.

Graham, S., & Harris, K. R. (2002). Prevention and intervention for struggling writers. In M. R. Shinn, H. M. Walker & G. Stoner (Eds.), *Interventions for academic and behavior problems II: Preventive and remedial approaches* (pp. 589-610). Bethesda, MD: NASP Publications.

Graham, S., Harris, K., & Hebert, M. A. (2011). *Informing writing: The benefits of formative assessment. A Carnegie Corporation Time to Act report.* Washington, DC: Alliance for Excellent Education.

Graham, S., Harris, K. R., & Larsen, L. (2001). Prevention and Intervention of Writing Difficulties for Students with Learning Disabilities. *Learning Disabilities Research & Practice, 16*, 74-.

Graham, S., & Hebert, M. A. (2010). Writing to read: Evidence for how writing can improve reading. A Carnegie Corporation Time to Act Report. Washington, DC: Alliance for Excellent Education.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445-476.

Graves, D. H. (1994). *A fresh look at writing*. Portsmouth, N.H: Irwin Publishing.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*, 499- 510.

Greenberg, K. L. (1992). Validity and reliability issues in the direct assessment of writing. *WPA: Writing Program Administration, 16*, 7-22.

Gronlund, N. E. & Linn, R. L. (1990). *Measurement and evaluation in teaching*. New York, NY: Macmillan.

Gunning, T. (2002). *Assessing and correcting reading and writing difficulties*. Boston: Allyn and Bacon.

Hammill, D. D., & Hresko, W. (1994). *Comprehensive scales of student abilities quantifying academic skills and school-related behavior through the use of teacher judgements*. Austin, TX: Pro-Ed.

Hammill, D. D., & Larsen, S. C. (1996). *TOWL-3: Test of written language*. Austin, TX: Pro-Ed.

Hammill, D. D., & Larsen, S. C. (2009). *TOWL-4: Test of written language*. Austin, TX: Pro-Ed.

Hammill, D. D., Pearson, N. A., & Wiederholt, J. L. (1996). *Comprehensive Test of Nonverbal Intelligence.* Austin, TX: Pro-Ed.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), Assessing second language writing in academic contexts (pp. 241–276). Norwood, NJ: Ablex.

Harris, K. R.. & Graham, S. (1996). *Making the writing process work: Strategies for composition and self-regulation*. Cambridge, MA: Brookline Books.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20,* 281-307. doi:10.1080/0969594X.2012.742422

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), The science of writing: Theories, methods, individual differences, and applications (pp. 1–27). Mahwah, NJ: Lawrence Erlbaum Associate.

Hessler, T., & Konrad, M. (2008). Using curriculum-based measurement to drive IEPs and instruction in written expression. *Teaching Exceptional Children, 41*, 28-37.

Hessler, T., Konrad, M., & Alber-Morgan, S. (2009). Assess student writing. *Intervention in School and Clinic, 45,* 68-71. doi:10.1177/1053451209338400

Holt, D. (1993). Holistic scoring in many disciplines. *College Teaching, 41,* 71-74.

Howell, D. C. (2002) *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Houck, C. K., & Billingsley, B. S. (1989). Written expression of students with and without learning disabilities: Differences across the grades. *Journal of Learning Disabilities, 22,* 561-572.

Hughes, C. A., & Dexter, D. D. (2011). Response to intervention: A research-based summary. *Theory into Practice, 50,* 4-11. doi:10.1080/00405841.2011.534909

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43,* 253-263. doi:10.3102/0013189X14542154

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60,* 237-263. doi:10.3102/00346543060002237

Huot, B. A. (2002). *(Re)articulating writing assessment for teaching and learning*. Logan: Utah State University Press.

Jeffery, J. V. (2009). Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing, 14,* 3-24. doi:10.1016/j.asw.2008.12.002

Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production- independent, and accurate-production scores. *School Psychology Review, 34,* 27-44.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How Can We Improve the Accuracy of Screening Instruments? *Learning Disabilities Research & Practice, 24,* 174-185.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2,* 130-144. doi:10.1016/j.edurev.2007.05.002

Juzwik, M. M., Curcic, S., Wolbers, K., Moxley, K. D., Dimling, L. M., & Shankland, R. K. (2006). Writing into the 21st century: An overview of research on writing, 1999 to 2004. *Written Communication, 23,* 451-476. doi:10.1177/0741088306291619

Kim, Y., Al Otaiba, S., Folsom, J. S., Greulich, L., & Puranik, C. (2014). Evaluating the dimensionality of first-grade written composition. *Journal of Speech, Language, and Hearing Research, 57,* 199-211.

Magrath, C. P., & Ackerman, A. (2003). The neglected ‘‘R”: The need for a writing revolution. The National Commission on Writing. New York, NY: College Entrance Examination Board.

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

Marston, D., Lowry, L., Deno, S., & Mirkin, P. (1981). *An analysis of learning trends in simple measures of reading, spelling, and written expression: A longitudinal study.* (Report Report No. 49). Minneapolis, MN: University of Minnesota, Institute for Research on Learning Disabilities.

McAlenney, A. L., & McCabe, P. P. (2012). Introduction to the role of curriculum-based measurement in response to intervention. *Reading Psychology, 33*, 1-7. doi:10.1080/02702711.2012.630599

McMaster, K. L., Xiaoqing Du., & Pétursdóttir, A. (2009). Technical features of curriculum-based measures for beginning writers. *Journal of Learning Disabilities, 42*, 41-60. doi:10.1177/0022219408326212

McMaster, K. L., Xiaoqing, D. U., Seungsoo, Y. E. O., Deno, S. L., Parker, D., & Ellis, T. (2011). Curriculum-based measures of beginning writing: Technical features of the slope. *Exceptional Children, 77*, 185-206.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing. *The Journal of Special Education, 41*(2), 68-84. doi:10.1177/00224669070410020301

McMillan, J. H. (2012). *Educational research: Fundamentals for the consumer* (6th ed.). Boston, MA: Pearson.

Medina, A. L. (2006). The parallel bar: Writing assessment and instruction. In J. S. Schumm (Ed.), *Reading assessment and instruction for all learners* (pp. 381-430). New York, NY: Guilford Press.

Miller, W. H. (1995). *Alternative assessment techniques for reading and writing.* San Francisco, CA: Jossey-Bass.

Miller, D. M., Linn, R. L., & & Gronlund, N. E. (2013). *Measurement and assessment in teaching* (11th ed.). Boston, MA: Pearson.

Moffett, J. *(*1968*).* Teaching the universe of discourse. *Boston, MA: Houghton Mifflin.*

Morris, L. (2013). *RTI meets writer's workshop: Tiered strategies for all levels of writers and every phase of writing*. Thousand Oaks, CA: Corwin.

Morrow-Howell, N. (1994). The M word: Multicollinearity in multiple regression. *Social Work Research, 18*, 247-251.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*, 5-12.

National Center on Response to Intervention (2010). *Essential components of RTI – A closer look at Response to Intervention.* Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention.

National Governors Association (2010). *Common core state standards initiative*. Washington, DC: National Governors Association.

National Governors Association. (2005). *Graduation counts: A report of the National Governors Association task force on state high school graduation data.* Washington, D.C.: National Governors Association.

National Writing Project. (2008). *An analysis of scoring systems employed in state writing assessment programs throughout the United States.* University of California, Berkeley; Berkeley, CA: National Writing Project.

Nunnally, J. (1967). *Psychometric theory.* New York, NY: McGraw-Hill.

Nystrand, M. (2006). The social and historical context for writing research. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 11-27). New York, NY: Guilford Press.

Nystrand, M. (1982). *What writers know: The language, process, and structure of written discourse*. New York: Academic Press.

Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal Of Educational Psychology*, *101*, 37-50. doi:10.1037/a0013462

Overton, T. (2009). *Assessing learners with special needs: An applied approach*. Upper Saddle River, NJ: Merrill.

Parker, D. C., McMaster, K. L., Medhanie, A., & Silberglitt, B. (2011). Modeling early writing growth with curriculum-based measures. *School Psychology Quarterly, 26*, 290-304. doi:10.1037/a0026833

Parker, R. I., Tindal, G., & Hasbrouck, J. (1991a). Countable Indices of Writing Quality: Their Suitability for Screening-Eligibility Decisions. *Exceptionality: a Research Journal, 2*, 1-17.

Parker, R. I., Tindal, G., & Hasbrouck, J. (1991b). Progress Monitoring with Objective Measures of Writing Performance for Students with Mild Disabilities. *Exceptional Children, 58,* 1, 61-73.

Penner-Williams, J., Smith, T. E. C., & Gartin, B. C. (2009). Written language expression: Assessment instruments and teacher tools. *Assessment for Effective Intervention, 34*, 162- 169. doi:10.1177/1534508408318805

Persky H., Daane, M., & Jin, Y. (2003). *The nation's report card: Writing, 2002*. Washington, DC: U.S. Department of Education.

Pierangelo, R. & Giuliani, G. A. (2012). *Assessment in special education: A practical approach* (4th ed.). Pearson: Boston.

Razali, N. M. & Wah, Y. B. (2011) Power comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2,* 21-33.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*, 18-39. doi:10.1016/j.asw.2010.01.003

Ritchey, K. D., & Coker, D. L. (2013). An investigation of the validity and utility of two curriculum-based measurement writing tasks. *Reading & Writing Quarterly, 29*, 89-119. doi:10.1080/10573569.2013.741957

Roberts, G., Wexler, J., Vaughn, S., Fall, A., Pyle, N., & Williams, J. (2012). Efficacy of an individualized reading intervention with secondary students. Evanston, IL: Society for Research on Educational Effectiveness.

Saddler, B., & Asaro-Saddler, K. (2013). Response to intervention in writing: A suggested framework for screening, intervention, and progress monitoring. *Reading & Writing Quarterly, 29*, 20-43. doi:10.1080/10573569.2013.741945

Salahu-Din, D., Persky , H., & Miller, J. (2008). *The nation's report card: Writing 2007* (NCES Publication No. 2008-468). Washington DC: National Center for Education Statistics.

Salkind, N. J. (2006). *Tests and measurement for people who (think they) hate tests and measurement.* Thousand Oaks, CA: Sage.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2010). *Assessment in special and inclusive education* (10th ed.). Belmont, CA: Wadsworth/Cencage Learning.

Sattler, J. M (2008). *Assessment of children: Cognitive foundations*. San Diego, CA: J.M. Sattler.

Shavelson, R., and Webb, N. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.

Sheils, M. "Why Johnny Can't Write." *Newsweek,* Dec. 8, 1975, pp. 58-63.

Shinn, M. R., & Bamonto, S. (1998). Advanced applications of curriculum-based measurement: "Big ideas" and avoiding confusion. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 1-31). New York, NY: Guilford Press.

Shriner, J. G. & Thurlow, M. L. (2012). Curriculum-based measurement, progress monitoring, and state assessments. In C. A Espin & K. L. McMaster (Eds.), *A measure of success: The influence of curriculum-based measurement on education* (pp. 247-260). Minneapolis, MN: University of Minnesota press.

Spandel, V. (2013). *Creating writers: 6 traits, process, workshop, and literature*. Boston, MA: Pearson.

Starch, D. & Elliot, E. (1912). Reliability in grading high school work in English. *School Review, 20*, 442-457.

Stone, E. (1974). *Donald writes no more: A biography of Donald Goines*. Los Angeles, CA: Holloway House.

Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics (4th ed.). New York, NY: Harper Collins College Publishers.

Thorndike, R. M. (1990). Would the real factors of the Stanford-Binet Fourth Edition please come forward? *Journal of Psychoeducational Assessment, 8*, 412-435.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education*. Boston: Prentice Hall.

Tindal, G., & Parker, R. (1989). Assessment of written expression for students in compensatory and special education programs. *The Journal of Special Education, 23*, 169-183. doi:10.1177/002246698902300204

Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research & Practice, 6*, 211-218.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Videen, J., Deno, S., & Marston, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression* (Research Report No. 84). Minneapolis, MN: University of Minnesota, Institute for Research on Learning Disabilities.

Wallace, T., Espin, C. A., McMaster, K., Deno, S. L., & Foegen, A. (2007). CBM progress monitoring within a standards-based system. *The Journal of Special Education, 41*(2), 66-67. doi:10.1177/00224669070410020201

Wayman M. M., Wallace T., Wiley H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading: A literature review. The Journal of Special Education, 41, 85-120.

Weissenburger, J. W., & Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *Journal of School Psychology, 43*, 153-169. doi:10.1016/j.jsp.2005.03.002

West, S. G., Finch, J. F. and Curran, P. J. (1995) Structural equation models with non-normal variables: Problems and remedies. In Hoyle, R. (Ed.), *Structural Equation Modeling: Concepts, Issues, and Applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Zimmerman, B. J., Bonner, S., & Kovach, R. (1996). *Developing self-regulated learners: Beyond achievement to self-efficacy.* Washington, DC: American Psychological Association.

APPENDIX A

Human Subjects Office/
Institutional Review Board (IRB)
105 Hardin Library for the Health Sciences
600 Newton Road
Iowa City, Iowa 52242-1098
319-335-6564 Fax 319-335-7310
irb@uiowa.edu
http://research.uiowa.edu/hso

May 26, 2015

TO: Paula Ganzeveld
Graduate College
John Hosp

FROM: John Wadsworth, PHD
IRB Chair or Chair Designee

RE: Not Human Subjects Research Determination

I have reviewed the information submitted with your project titled 201505793 Combining Quality and Curriculum-Based Measurement: A Suggested Assessment Protocol in Writing. I have determined that the project described in the application *does not* meet the regulatory definition of human subjects research and does not require review by the IRB, because the project is to increase the validity of a measure and is not about humans behavior.

We appreciate your care in submitting this application to the IRB for review. If the parameters outlined within this Human Subjects Research application request change, re review and/or subsequent IRB review may be required.

Please don't hesitate to contact me if you have any questions. The Human Subjects Office can be reached via phone (319)-335-6564 or email irb@uiowa.edu.

APPENDIX B

CPU – SECOND GRADE WRITING RUBRIC					
STUDENT:				DATE:	
INDICATORS	0	1	2	3	3+
CONTENT/IDEAS: RESPOND TO ALL PARTS OF THE PROMPT INCLUDING A TOPIC AND IDEAS OF ANSWERING QUESTIONS OF WHO?, WHAT?, WHEN?, HOW?, WHERE, AND WHY?					
CONTENT	Comments:	Meets SOME requirements of prompt Main idea of a topic may not be clear and insufficient details, facts, and definitions	Meets MOST requirements of prompt Main idea of a topic is named and some details, facts, definitiions included	Meets ALL requirements of the prompt Main idea of topic is introduced and communicated clearly using details, facts, definitions	Comments:
INDICATORS	0	1	2	3	3+
ORGANIZATION: SENTENCE STRUCTURE AND BEGINNING, MIDDLE, END					
ORGANIZATION	Comments:	Abrupt beginning/ end; body not developed Provides few supporting details	A recognizable introduction, body, conclusion Provides some supporting details	Clear introduction, body, conclusion Provides specific supporting details	Comments:
INDICATORS	0	1	2	3	3+
WORD CHOICE: CHOICE OF WORDS PAINTS A PICTURE FOR THE READER. HAS NOUNS, VERBS, DESCRIBING WORDS, AND SHOWS FEELINGS					
WORD CHOICE	Comments:	Rambling or choppy sentences, run-ons Many sentences begin the same way; little variance in structure	Generally fluent, with occasional choppiness Some varieties in sentence beginnings, structures, lengths	Fluent, easy to read Effective variety of sentence beginnings, structures, lengths	Comments:

9 MEETS STANDARDS
6-8 DEVELOPING STANDARDS
0-5 APPROACHING STANDARDS