The Pennsylvania State University

The Graduate School

College of Education

IDENTIFYING CBM WRITING INDICES FOR EIGHTH GRADE STUDENTS

A Dissertation in

School Psychology

by

Janelle Amato

Submitted in Parital Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2008

The Dissertation of Janelle Amato was reviewed and approved* by the following:

Marley W. Watkins
Professor of Education
Thesis Advisor
Chair of Committee

Barbara A. Schaefer
Associate Professor of Education

Richard M. Kubina
Associate Professor of Education

Liza M. Conyers
Associate Professor of Education

James C. DiPerna
Associate Professor of Education
Professor-in-Charge of School Psychology

*Signatures are on file in the Graduate School.

Abstract

Curriculum-based measurement (CBM) is an alternative to traditional assessment techniques. Initially CBM was developed so that teachers could monitor elementary students' growth in basic skill areas. Thus, most of the CBM research was conducted at the elementary school level. Recently, educators have become interested in applying CBM at the secondary level. Therefore, technical work has begun in order to identify CBM writing indices that are psychometrically sound for purposes of monitoring older students' growth. CBM is thought to be a valuable assessment tool in writing due to the ease with which educators can produce and score CBM probes. Additionally, CBM is a direct assessment of written expression. That is, CBM requires students to formulate a written product. This is in contrast to most traditional measures. The present study examined the predictive validity of the following CBM writing indices: total words written, words spelled correctly, percentage of words spelled correctly, number of correct word sequences, percentage of correct word sequences, number of correct minus incorrect word sequences, number of sentences, number of correct capitalization, total punctuation marks, and number of correct punctuation marks. Regression analyses revealed simple fluency measures are not sufficient for assessing secondary students' writing ability. Results indicated that a more complex fluency measure, the number of correct punctuation marks, and an accuracy-based measure, the percentage of correct word sequences, were the best predictors of a well-designed written expression test for eighth grade students. However, overall results of the current study did not lend strong support to the use of CBM to assess and monitor writing skill at the secondary level.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my gratitude to my thesis adviser, Dr. Marley Watkins, for his endless patience, support, and attention to detail. Additionally, I would like to thank Dr. Barbara Schaefer, Dr. Rick Kubina, and Dr. Liza Conyers for their guidance and assistance in completing this project.

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

I am particularly grateful to my husband, Antonio, for his unending assistance, insight, and support, and to my children, Bella and Luca, whose love and infectious laughter got me through the last year. I would also like to acknowledge my biggest fans, Mom, Dad, Stacy, Georgie, and David, for being my constant inspiration and my loudest cheerleaders.

Finally, I am grateful to all of my family and friends (there are too many to mention) who encouraged me to complete this project. Without the support and confidence of these incredible people, this daunting task would have been impossible.

Writing is an essential skill for academic success, professional advancement, and personal communication. Writing, along with speaking, is a primary mechanism for verbal communication. Students are often required to communicate their knowledge through their writing. As students progress through grade levels, their need for writing skills increases as they are required to write more papers and to write more on tests (Bradley-Johnson & Lesiak, 1989). Because writing skills are essential for academic success, many children who experience difficulty in writing are at serious risk for school failure (Reschly, 1992). The latest results of the *Nation's Report Card* (U.S. Department of Education, Institute of Education Sciences, & National Center for Educational Statistics, 2003) indicated that 14%, 15%, and 26% of fourth, eighth, and twelfth grade students, respectively, were not able to write at the basic level. Moreover, the number of twelfth grade students unable to write at the most basic level has significantly increased since 1998.

Many students who experience difficulties with written expression in school subsequently display deficits in writing skills as adults (Gajar, 1989). Adults with writing disorders often experience frustration at work due to the high demand of writing in many jobs. In a 1988 survey, 80% of employers stated that the ability to write accurate messages was the most important skill an employee would need and 67% of the employers listed the ability to write requests as the second most important skill (Algozzine, O'Shea, Stoddard, & Crews, 1988). Yet employers have observed that many college graduates cannot write and have weak thinking skills (Thomas, 1994).

In recognition of the importance of written expression skills, writing has always been an important part of the school's curriculum (U.S. DOE, IES, & NCES, 2003).

However, it has received less attention from educators and researchers than reading and math. Early research on writing skills tended to focus primarily on spelling, until written expression was included as a category of learning disabilities in Public Law 94-142 (Bradley-Johnson & Lesiak, 1989). In the past two decades, an increased number of research studies in the area of written expression have been conducted.

RESEARCH ON WRITTEN EXPRESSION

Efforts to improve written language instruction in the schools have been accompanied by an increased number of research studies examining different aspects of written language. As with all concepts, how the terms are defined have a significant impact on how the area is researched. Given the complexity of the writing process, researchers have struggled with establishing a clear definition or model of written expression (Cole, Haley, & Muenz, 1997; Hooper et al., 1994). Hooper and colleagues defined writing "as a problem-solving process whereby authors attempt to produce visible, understandable, and legible language reflecting their declarative knowledge (i.e., knowledge of the topic)" (p. 377). Despite the ongoing efforts of research to increase our understanding of the writing process, there have been few models developed to explain or examine the process of writing. The following models are most frequently encountered in the literature and therefore warrant elaboration.

Hayes and Flower Model

Hayes and Flower (1980) proposed the most influential model of the cognitive processes involved in writing (Berninger et al., 1997; Hooper et al., 1994). Hayes and Flower identified three main components that interact to affect writing: the writing process, the task environment, and the writer's long-term memory (Albin, Benton, &

Khramtsova, 1996). The main features of this model lie within the writing process and include planning, translating, and reviewing (Albin et al.). Planning involves what to say, translating helps transform those plans into written text, and reviewing is a process in which writers continually evaluate and revise text according to standards and expectations (Albin et al.; McCutchen, Covill, Hoyne, & Mildes, 1994). These three processes are interrelated and recursive and writers do not proceed in a rigid order from one process to the next.

Along with the writing process, the task environment and the writer's long-term memory can affect written expression. According to Hayes and Flower (1980), the task environment includes aspects such as the topic about which the author is writing, the audience for which the author is writing, and motivational factors. The task environment also includes text that is already produced by the writer. For example, many authors create notes and outlines on the topic to help develop their written product (Albin et al., 1996; Hooper et al., 1994). Long-term memory, the last component of the Hayes and Flower's writing model, is where writers store information regarding the different purposes of writing (e.g., narrative, informative, persuasive). It is also in long-term memory where authors store information regarding the topic and audience. It is known that students' background knowledge and prior experiences tend to influence the topic about which they choose to write (Albin et al.).

Berninger Model

More recently, Berninger and colleagues (Berninger, 2000; Berninger et al., 1997; Berninger et al., 2002; Berninger & Graham, 1998; Berninger, Whitaker, Feng, Swanson, & Abbot, 1996) examined aspects of written language such as handwriting skills,

spelling, and written composition. Berninger's view of writing can be figuratively represented by a triangle. The vertices of the base represent transcription skills and self-regulation skills, and the acme represents the goal of text generation (Berninger, 2000; Berninger et al., 2002). Transcription occurs when the writer translates language representations from working memory into orthographic symbols on paper or on a computer screen. Self-regulation skills guide the writing process and include higher level processes such as goal-setting, monitoring, and reviewing and revising text. The more automatic the low level processes (e.g., transcription skills), the more resources are available for higher level processes (e.g., self-regulating skills; Berninger et al., 1996; Berninger et al., 1997). Berninger's model of the writing process is simply viewed as transcription and self-regulation skills working together to reach the goal of text generation.

Differences between Well-Written and Poorly-Written Essays

Although there are not many well-defined models of writing, there are many studies that describe the characteristics of essays that were written by skilled and unskilled writers, and the processes used by skilled and unskilled writers to create their final product (Cole, Haley, et al., 1997; Graham, 1990; Graham & Harris, 1997; Hooper et al., 1994; MacArthur & Graham, 1987). Although not a theoretical model of the writing process, these distinctions help define good writing. The characteristics of the final product and of the processes used are important not only in distinguishing between skilled and unskilled writers, but also for developing treatment and instructional programs (Hooper et al.).

In general, skilled writers are described as goal-directed learners who apply various self-regulation strategies such as planning, revising, organizing, monitoring, and evaluating (Cole, Haley, et al., 1997; Graham & Harris, 1997). They understand the goals of the writing assignment and have a greater knowledge about their writing topic and their audience (Hooper et al., 1994). In addition, skilled writers generate more ideas and eliminate their less productive ideas as they revise and edit. These characteristics add to the smoothness and cohesiveness of the writer's final product (Graham & Harris; Hooper et al.).

In contrast, children who find writing challenging use less sophisticated approaches to writing. They not only display deficits in their use of self-regulation strategies, but also show problems in generating text ideas. Unskilled writers are less likely to revise their spelling, punctuation, grammar, or text ideas, resulting in poorly written text (Graham & Harris, 1997; Hooper et al., 1994). Additionally, students who experience difficulties in writing tend to have shorter compositions and provide the audience with little detail or elaboration when compared to skilled writers.

Graham and Harris (1997) provided three possible reasons why unskilled writers produce shorter essays. The first reason may be because students who struggle with writing terminate their writing process too soon. A study by Graham (1990) provided some evidence for this proposition. Fourth and sixth grade students with learning disabilities generally wrote for 6 or 7 minutes when writing an essay. However, when verbally prompted to write more, these students generated substantial increases in the amount of text written.

Graham and Harris (1997) also suggested that unskilled writers may produce shorter essays when compared to skilled writers due to poorly developed mechanical skills. In a study by MacArthur and Graham (1987), fifth- and sixth-grade students with learning disabilities produced longer stories and improved the quality of their stories when they dictated their essays versus handwriting them or typing them on a word processor. These results are consistent with Graham (1990) who examined the effects of mechanical skills on writing for fourth- and sixth-grade students with learning disabilities. Graham also found that under normal conditions, students generated better quality essays when the stories were dictated rather than written. However, unlike McArthur and Graham, Graham did not observe an increase in the length of output as a result of mode.

Finally, the third possible reason provided by Graham and Harris (1997) is not related to writing skills, but related to topic knowledge and interest. Graham and Harris proposed that students who lack knowledge or interest in a topic will generate less text than those students who are knowledgeable and interested in the topic. In the past, researchers had largely theorized that individual interest influences students' writing; however, these claims are not supported by research (Hidi & McLaren, 1990; Hidi & McLaren, 1991). For example, Hidi and McLaren (1991) found that students in the fourth and sixth grades did not write longer or qualitatively better essays on topics that they identified as interesting when compared with topics that they identified as uninteresting. In contrast, other research has supported the hypothesis that topic knowledge does influence length and quality of a written product (DeGroff, 1987; Kellogg, 1987; McCutchen, 1986). For example, McCutchen found that high school students who were

knowledgeable about football wrote lengthier and more coherent texts than did students who had low levels of knowledge about football.

ASSESSING WRITING SKILLS

Given the absence of an adequate model of written language, assessment methods for written expression have lagged behind other academic domains. Without a clear definition or understanding of the writing process, it becomes challenging to produce a comprehensive assessment of writing that leads to accurate identification of difficulties and provision of effective instruction (Hooper et al., 1994). Although there are obstacles involved with developing writing assessments, valid assessment methods are necessary for four reasons. First, the most common purpose of assessing writing is to determine the presence or absence of a writing disorder and subsequent eligibility for special educational services. Second, students' strengths and weaknesses should be targeted in order to develop and implement appropriate instruction. Third, valid assessment procedures are also valuable for monitoring student progress and determining treatment effectiveness. Finally, assessment methods can provide students with direct feedback, which has been shown to improve academic skills (Burns & Symington, 2003; Hooper et al., 1994; Shapiro, 1989).

Writing assessments should measure a variety of skills because students with writing problems tend to have a wide range of skill deficits (Bradley-Johnson & Lesiak, 1989; Tindal & Parker, 1989). Problems can stem from the mechanical aspects of writing, such as handwriting, spelling, punctuation, or capitalization; to the content aspects of writing, such as the cohesiveness, organization, and quality of the product (Li & Hamel, 2003). In the past decade, research regarding the above areas have multiplied, which has

resulted in an increase in the number and quality of assessment instruments (Bradley-Johnson & Lesiak). Nonetheless, many written assessment methods are still psychometrically problematic and time-consuming to produce, administer, or score (Watkinson & Lee, 1992). There are a variety of assessment methods (e.g., indirect or direct) and scoring options (e.g., subjective or objective) that are discussed in the literature. A review of these methods follows.

Indirect Assessment

Indirect assessment of written language does not require students to formulate and produce an essay or a composition. Instead, the primary purpose of indirect measures is for students to demonstrate knowledge of writing conventions. For example, many indirect measures of writing assess students' abilities to detect and correct grammatical errors, detect and correct spelling mistakes, or revise a sentence or paragraph (Hooper et al., 1994; Watkinson & Lee, 1992).

Indirect measures are often norm-referenced and have the advantage of being relatively objective. Because indirect measures are objective, the scoring procedures for these tests are generally simple and result in high interscorer reliability when compared to direct assessments of writing. Additionally, indirect measures of written expression are generally representative of the domain and include a large number of items (Shinn, 1989). Thus, indirect measures can be useful for screening and eligibility decisions regarding special education (Watkinson & Lee, 1992; Shinn, 1989).

Although indirect measures are valuable, they are problematic for educational decision making (Salvia & Ysseldyke, 2001; Shinn, 1989). For example, indirect measures of writing frequently do not match a school's curriculum, nor do they resemble

typical classroom writing assignments (Salvia & Ysseldyke; Watkinson & Lee, 1992). The scores earned on indirect measures are also limited in the amount and quality of information they provide. Thus, educators cannot infer that a low score on an indirect measure indicates a student's failure to profit from instruction (Salvia & Ysseldyke).

Indirect measures such as proofing and dictation are often included in standardized, norm-referenced tests (Watkinson & Lee, 1992). Unfortunately, these types of measures are not suitable for planning daily instruction (Salvia & Ysseldyke, 2001; Shinn, 1989). Standardized, norm-referenced tests are designed to produce stable scores and are not sensitive to small changes in student performance. Thus, students who are developing their writing skills slowly may show no gains when their progress is tested with standardized, norm-referenced measures. Moreover, indirect measures generally provide teachers with only one global score from the test. They don't provide the opportunity to analyze errors or examine strengths and weaknesses revealed by the student's performance. Therefore, scores from indirect measures have little validity for evaluating student progress (Salvia & Ysseldyke; Shapiro, 1987; Shinn, 1989; Tindal & Parker, 1989).

Direct Assessment

In contrast to indirect assessment, direct assessment of written language requires students to formulate and produce an essay or a composition. Thus, direct measures of written expression produce scores that have much stronger content validity than indirect measures of written expression (Hooper et al., 1994; Tindal & Parker, 1989; Watkinson & Lee, 1992). Direct assessment allows the examiner to observe the students' writing process, as well as evaluate what the students were taught in class (Watkinson & Lee).

Although direct assessment methods do not require sophisticated test construction skills, they are more time consuming to score than are indirect methods. Additionally, many of the scoring rubrics for direct assessment require intensive training to obtain acceptable levels of reliability (Hooper et al.).

Direct assessment methods are usually divided into two types: process methods and product methods. Process methods are dynamic and involve determining the problem-solving steps or self-regulating strategies that students use during a written expression task. This could be done while the student is writing (e.g., think-aloud protocols) or after the student completes the writing task (e.g., semi-structured interviews) (Hooper et al., 1994). Although these procedures may provide valuable formative information (Hooper et al.), it is generally the written product that is assessed (Isaacson, 1988; Jones, 1998). Frequently, if the final product is deemed good, process strategies are judged to be successful (Isaacson). Although examiners may use work samples that students have already produced, generally students are asked to generate a written product specifically for evaluation purposes. This helps control for other factors such as misunderstanding of a task or lack of interest (Hooper et al.). Once the final product is completed, the examiner needs to choose the most appropriate scoring method.

Subjective Scoring

The products of direct assessment of written expression are assessed using two scoring approaches: subjective and objective. There are three types of subjective scoring that frequently appear in the literature: holistic, analytic, and primary trait.

Holistic

Holistic scoring is generally an overall impression of the final product represented by a single score. This single score represents all aspects of the writing taken together (Dahl & Farnan, 1998; Isaacson, 1999; Tindal & Parker, 1989). Holistic scoring is most useful for placement or grading; however, it does not contribute to treatment planning because it does not identify the writer's weaknesses or strengths (Dahl & Farnan; Hooper et al., 1994; Isaacson).

Assessment methods are only valuable when the scores they produce are both valid and reliable. Scores produced by subjective methods possess a high level of face validity, but have questionable construct validity and reliability (Hooper et al., 1994). For example, Charney (1984) reported that holistic scores could easily be influenced by the length of the essay, the number of errors, and the neatness of the written composition. However, it has been reported that the validity and reliability of holistic scores could improve if specific grading criteria are defined. In addition, the evaluators using the criteria should be rigorously trained until an acceptable rate of reliability is obtained. Thus, valid and reliable holistic scores are dependent on detailed scoring criteria, appropriate training, and fidelity of implementation of scoring criteria (Bradley-Johnson & Lesiak, 1989; Hooper et al.). The importance of training to the reliability of holistic scores has long been known. For example, in 1934 Stalnaker (as cited in Cooper, 1977) reported an increase of a reliability range from .30 to .75 prior to training to a range of .73 to .98 after training. Thus, it appears that given specific criteria, holistic scoring can be a reliable and an efficient scoring method to use, typically taking 1 to 2 minutes per paper (Bradley-Johnson & Lesiak, 1989).

Analytic

Although holistic scoring can be valid, reliable, and efficient, it still offers little information for treatment planning. Therefore, analytic scoring may be more useful for assessing the writer's strengths and weaknesses, as it provides several scores rather than just one holistic score. Analytic scoring separately analyzes different characteristics of writing providing detailed information about a variety of different writing skills (Dahl & Farnan, 1998; Hooper et al., 1994; Isaacson, 1999; Tindal & Parker, 1989). For example, Spandel and Stiggins (1997) suggested six characteristics of writing that can be analytically scored: ideas, organization, voice, word choice, sentence fluency, and conventions. Thus, unlike holistic scoring in which the student would receive one overall score, in analytic scoring the student would receive six scores for each characteristic of writing. Of the three types of subjective scoring methods, analytic scoring provides the most information to teachers for instruction and intervention purposes. However, the advantage of using analytic scoring methods is dependent on the number of writing components chosen and the weight given to each (Hooper et al.).

Primary Trait

Similar to holistic scoring, primary trait scoring methods yield one score; however, this score represents a specific purpose of the writing task. For example, students may be asked to arrange the events of a story in a particular order and the score they receive reflects their performance on this particular task. Or they may have been assigned to write for a specific audience and are then evaluated on how well they accomplished this assignment (Bradley-Johnson & Lesiak, 1989; Dahl & Farnan, 1998; Tindal & Parker, 1989). This type of scoring may be useful for teachers to identify

whether the primary purpose of a given writing task has been met (Dahl & Farnan). In addition, like holistic and analytic scoring methods, a primary trait score can be psychometrically sound if the proper amount and type of training is provided to examiners (Cooper, 1977). However, primary trait scoring is not useful for large-scale assessment, as it only focuses on one trait (Dahl & Farnan).

Objective Scoring

The second major method used to score students' writing protocols involves countable indices. In contrast to the subjective methods just described, these scoring methods require the examiner to count a given dimension of the final product (e.g., total number of words written) and therefore are objective in nature (Tindal & Parker, 1989). The most common type of direct assessment that uses objective scoring is called curriculum-based measurement (CBM).

Curriculum-based Measurement

In response to the limitations of standardized, norm-referenced tests, CBM was originally developed in the early 1970s by Stanley Deno to provide teachers with an accurate and effective method to monitor instruction. By monitoring instruction, teachers would be better able to assist students in improving academic performance (Shinn & Bamonto, 1998). Although CBM was primarily developed as an assessment method for special education teachers to evaluate the effectiveness of instruction, it is now employed for a variety of other purposes. These include, but are not limited to, screening at-risk students, identifying a problem, monitoring student growth, and recommending and evaluating inclusion (Deno, 2003).

CBM scores are indicators of student performance primarily in the area of basic skills. Therefore, CBM scores are not direct measures of the underlying construct (Shinn & Bamonto, 1998). Due to CBM's limited sampling of behaviors, it has received criticism from researchers regarding its curricular validity. For example, Mehrens and Clarizio (1993) stated that a local curriculum is much broader than the skills sampled by CBM; thus, CBM scores cannot be a valid measurement of student achievement. Shinn and Bamonto's responded to this criticism by asserting that CBM was not intended to assess all skills in the curriculum, but rather was intended to index only basic skills.

The nature of CBM makes it a useful formative evaluation tool. Formative evaluation entails ongoing assessment and instructional modifications based on empirical data. CBM is a type of formative evaluation in that student progress can be measured and evaluated frequently using curricular materials. Decisions to modify instruction can then be based on the CBM data collected (Shinn & Bamonto, 1998). However, Mehrens and Clarizio (1993) asserted that CBM is limited in its use for designing or modifying interventions. They criticized CBM for its ineffectiveness in providing information on how to change instruction and stated that it was only useful in deciding when change was needed.

In order to use CBM as a formative evaluation tool, it must be standardized and able to be frequently administered (Shinn & Bamonto, 1998). Standardization of administration, scoring, and interpretation is needed before change in students' CBM scores can be attributed to learning. CBM also allows student performance to be assessed repeatedly across time. That is, this frequent measurement is possible due to the brevity of CBM probes, ranging from 1 to 5 minutes in length, and the ease with which educators

can produce and score CBM probes. Adding to its efficiency, professionals, educators, and parents can quickly learn to use CBM procedures in such a way that the data are reliable and valid (Deno, 2003; Shinn & Bamonto; Shinn, Rosenfield, & Knutson, 1989).

Moreover, CBM was designed to be a technically adequate method for collecting information regarding a particular academic area. Multiple studies have documented the reliability and validity of CBM scores when CBM is used to measure the basic skill areas of reading, mathematics, and writing (see Marston [1989] and Good & Jefferson [1998] for reviews). Although reliability and validity evidence has been documented in all three areas, most of the CBM literature has focused on reading and math.

In CBM reading, students are typically asked to read aloud for 1 minute while the examiner tracks the words that are read correctly and incorrectly. The most common index obtained from a CBM reading probe is an oral reading fluency (ORF) score, which represents the number of words read correctly per minute. In CBM math, students compute answers to addition, subtraction, multiplication, and division problems for 2 to 5 minutes. The correct number of digits per minute is the primary index used (Shinn & Bamonto, 1998).

The criterion-related evidence reported in Marston's (1989) review also included students across grades and age levels. In the reading and math studies that Marston reviewed, a published norm-referenced test was generally used as the criterion. Correlation coefficients between students' oral reading fluency scores and different criterion test scores ranged from .63 to .90, with most coefficients above .80. In a more recent summary by Good and Jefferson (1998), the reported median validity coefficients for oral reading fluency scores and criterion test scores were slightly smaller, ranging

from .62 to .73. As preferred, Good and Jefferson only included studies that reported within-grade validity coefficients. Thus, these may be more accurate than the validity coefficients reported by Marston. Additionally, in a recent synthesis of the literature in CBM in reading Wayman, Wallace, Wiley, Ticha, and Espin (2007) concluded that oral reading fluency was found to be a good indicator of reading comprehension and not just a speed-of-processing measure.

In addition to criterion-related evidence, Marston (1989) also summarized the published reliability evidence of CBM scores through 1989. In general, the test-retest, parallel form, and interrater reliability evidence for reading and math probes were high. The reliabilities for scores derived from reading probes and specific math probes (i.e., addition, subtraction, and mixed math probes) ranged from .70 to .99. However, reliability of multiplication and division probes ranged from .48 to .95. Although most of the reliabilities reported were acceptable, they should be viewed with caution (Mehrens & Clarizio, 1993; Good & Jefferson, 1998). Many of the reliability studies included students across grades and age levels, thereby inflating reliability estimates.

In contrast to the studies looking at validity evidence for CBM scores in reading, the studies in the area of math have consistently yielded lower validity coefficients. In Marston's (1989) review, the validity coefficients for CBM math scores and criterion tests scores ranged from .26 to .67, with few correlation coefficients exceeding .60. In Good and Jefferson's (1998) review, the median validity coefficient for 12 studies conducted in the third and fifth grades were .32 and .57, respectively. More recently, Foegan, Jiban, and Deno's (2007) review of the CBM literature in math revealed criterion-related validity coefficients between CBM math probes and criterion measures

ranged from .50 to .70. The consistently lower criterion-related correlations in math have led supporters of CBM to question the validity of published, norm-referenced math test scores. For example, Foegan et al. pointed out that the range of criterion-related validity coefficients for commercially available published math tests are in the same range as CBM math probes. However, this conclusion may be premature given the lack of evidence to support the utility of CBM over that of norm-referenced math tests.

Unlike the mixed research regarding the criterion-validity evidence for math CBM scores, Good and Jefferson (1998) consistently found that CBM scores in both reading and math differentiated between groups with diverse educational needs. This evidence added support for the use of CBM as a method for screening, referral, and assessment decisions. Although there is research supporting the use of CBM scores for distinguishing among students with different characteristics, there is little evidence that supports the ability of CBM scores to do so as well or better than published norm-referenced tests. Marston, Mirkin, and Deno (1984) reported that decisions regarding special educational services using CBM were more accurate than using teacher judgment. However, there was no evidence suggesting that CBM measures were more useful than published norm-referenced tests for referral purposes.

Less attention has been paid to the reliability and validity of CBM written expression scores (Fewster & Macmillan, 2002; Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002). Gansle and colleagues suggested that the complexities associated with assessing written language may be the reason why less research has been conducted in the area of CBM writing (e.g., lack of consensus regarding a definition or model of the writing process). In addition, the evidence supporting the reliability and validity of CBM

writing scores is less convincing than the evidence supporting the psychometric properties of reading and math CBM scores. However, past research has provided modest support for the reliability and validity of CBM scores in written expression, thus warranting further research on the usefulness of CBM writing probes as a tool that can provide reliable and valid scores (Marston; Good & Jefferson).

Research on CBM in Written Expression

Curriculum-based measures of written expression can be administered to students individually or in groups (Shapiro, 1989; Shinn, 1989). Typically, students are instructed to write a short story in response to a story starter or topic sentence. It is recommended that story starters and topic sentences be relevant to children to aid them in generating ideas and stories. Once the evaluator provides the students with a written copy of the story starter and orally reads the starting sentence, students are given 1 minute to plan their story. After 1 minute, students have 3 minutes to write and are encouraged by the evaluator to write until time is up.

Most of the research investigating CBM in written expression has been conducted with elementary school students. In addition, research has mainly focused on the relationship between different CBM scoring indices and traditional measures of written expression or holistic scoring. For the purpose of this study, the 10 written expression scoring indices most commonly used in the literature will be reviewed separately.

Total words written. Almost all CBM writing studies have examined the total number of words written. According to Shinn (1989), the total number of words written in the time period allotted, including words spelled incorrectly, can be used as an index of writing proficiency. If the student rewrites the story starter or writes a title, this is

included in the total words written. However, numerical representations and symbols are not included in this total. Research has consistently demonstrated that students who struggled with writing tended to produce shorter essays or compositions than skilled writers (Deno, Marston, & Mirkin, 1982; Graham & Harris, 1997; Houck & Billingsley, 1989; Poteet, 1979). For example, in a study comparing the writing of 48 students with learning disabilities and 48 normally achieving students in Grades 4, 8, and 11, on average, students with learning disabilities wrote fewer words (*m* = 140.9 words) than students who were achieving at an average level (*m* = 112.3 words; Houck & Billingsley).

Early studies also revealed significant correlations between the number of words written and other measures of writing skill (Anderson, 1982; Daiute, 1986; Deno, Marston, & Mirkin, 1982). One of the first studies examining the validity of total words written was conducted by Deno, Marston, and Mirkin (1982). Total words written, along with other scoring metrics, was correlated with nine criterion variables including selected subtests from the Test of Written Language (TOWL; Hammill & Larsen, 1978), the Stanford Achievement Test (Madden, Gardner, Rudman, Karlsen, & Mervin, 1978), and the Developmental Sentence Scoring System (Lee & Canter, 1971). With a sample of 136 elementary students in Grades 3 to 6, Deno, Marston, and Mirkin obtained moderate to high correlations, ranging from .58 for the Thematic Maturity subtest of the TOWL to .84 for the Developmental Sentence Scoring System, which was a measure of syntactic maturity.

Other studies investigated total words written as an indicator of change in student performance either following intervention or due to development. Marston, Lowry, Deno,

and Mirkin (1981, as cited in Espin, Shin, Deno, Skare, Robinson, & Benner, 2000) measured the sensitivity of total words written over time and across grades. The CBM writing measures were collected from 58 students in Grades 1 to 6 in the fall, winter, and spring. Results indicated that total words written consistently increased across time and across grade levels. These results were replicated by Deno, Marston, Mirkin, Lowry, Sindelar, and Jenkins (1982, as cited in Espin et al., 2000) in a large-scale study with 566 elementary students from three states. CBM writing probes were collected once in the fall and once in the spring. Total words written appropriately detected developmental changes from grade to grade and from fall to spring within grade levels.

A more recent study examined total words written as indicator of change following intervention in student performance for 45 students in Grades 3 and 4 (Gansle et al., 2004). This study examined changes in student performance on six indices following a brief intervention for writing. The intervention was based on self-regulated strategy development and consisted of brain-storming, note-taking, elaboration on topic ideas, planning, and constructive feedback.

Students were first required to complete a CBM writing probe following procedures outlined by Shinn (1989). After the first CBM writing probe was completed, the students were provided with a second story starter. However, this time the teacher brain-stormed with the students by having students suggest ideas for the story. Each student contributed at least one idea. The students were required to take notes and the instructor verified that all students complied. Additionally, the students practiced writing a complete sentence and, when necessary, the teacher helped the students. After this brief intervention was completed, the students were provided with a 3-minute break. After the

break, another story starter was provided and students completed the CBM writing task as outlined by Shinn.

Effects of the brief intervention were examined by comparing pre-intervention and post-intervention scores using t-tests. The authors did not expect large gains from pre- to post-intervention due to the brevity of the intervention. Therefore, the alpha level was set at .05 for each analysis. Although all six indices increased from pre- to post-intervention, a statistically significant effect was only found for total words written.

Although total words written is viewed as an appropriate index for overall writing ability for elementary students (Espin et al., 2000; Espin, Scierka, Skare, & Halverson, 1999), its appropriateness for older students has been questioned. Tindal and Parker (1989) were the first to investigate if the writing indices used for elementary students would be supported for use with older students. Tindal and Parker (1989) collected 6-minute writing samples from 172 students in Grades 6 to 8. When correlated with holistic scores of the same writing samples, total words written provided the lowest correlation coefficient among all the objective indices (r = .10). In addition, the number of words written did not statistically discriminate students in special education from students in remedial programs.

Other studies have also demonstrated the weakness of total words written as an acceptable scoring metric for CBM writing probes among secondary students (Espin et al., 2000; Espin, et al., 1999; Gansle et al., 2002; Malecki & Jewell, 2003; Parker, Tindal, & Hasbrouck, 1991; Watkinson & Lee, 1992; Weissenburger & Espin, 2005). For example, Espin and colleagues (2000) examined 3-minute and 5-minute writing probes in a sample of 112 students from Grades 7 and 8. The students were required to write two

story writing samples and two descriptive writing samples. The number of words written were correlated with teacher ratings of students' writing proficiency. One teacher rated the students' writing samples from 1 (*not yet proficient*) to 4 (*highly proficient*) on three main categories: (a) purpose, tone, and voice; (b) main ideas, details, and organization; and (c) sentence structure, mechanics, and legibility. Additionally, for the eighth-grade students the number of words written were correlated with scores from a district writing test that consisted of students writing a response to a given scenario. The district-wide writing samples were scored on a scale of 1 (*poor writing*) to 4 (*excellent writing*) by trained raters and the average interrater reliability was .73. Espin et al. reported that the alternate-form reliability coefficients for the 3-minute and 5-minute samples for total words written ranged from .73 to .77. However, they found that the criterion validity coefficients between words written and teacher ratings and the district-wide test ranged from .34 to .47. As with other research studies, total words written was low to moderately correlated with other measures of middle-school students' general writing performance.

More recently, Weissenburger and Espin (2005) examined the technical adequacy of total words written across three grade levels. They collected two 10-minute writing samples from 484 fourth, eighth, and tenth grade students and had them make a slash mark at the end of 3- and 5- minute intervals. Twenty writing samples from each grade level were randomly selected and double scored by a trained scorer. The mean interrater agreement percentage for total words written and two other CBM indices was 94.52% across all grade levels. Criterion-related validity coefficients were collected between total words written and a statewide assessment of achievement, The Wisconsin Knowledge and Concept Examinations (WKCE; CTB MacMillian/McGraw-Hill, 1993;

CTB/McGraw-Hill, 1996). The WKCE measures students' academic performance in several areas, including language arts. Weissenburger and Espin reported low to moderate correlation coefficients between total words written and the language arts subtests of the WKCE (r = .04 to .45). Although the correlations were significant ($p <$ .001) at the fourth-grade level, they were not significant at the eighth- and tenth-grade levels. In addition to the above criterion-related validity coefficients, Weissenburger and Espin correlated total words written with holistic scores collected from the direct writing assessment portion of the WKCE. Only the fourth and eighth graders completed this portion of the WKCE. All criterion-related validity coefficients were significant and ranged from .33 to .48. The correlation coefficients were higher for fourth graders than eighth graders in all three writing samples (3-, 5-, and 10-minute samples). Thus, Weissenburger and Espin concluded that the technical adequacy of TWW decreased with age.

In summary, most of the research at the elementary level has supported words written as a valid estimate of general writing ability. However, total words written was not a valid index of older students' written expression skills. Thus, measurement of students' writing ability at the secondary level may be more complex than at the elementary level.

Words spelled correctly. Correct spelling is an important component of writing because readers of the written product must be able to read the words written by the student. Several studies have compared the spelling skills of students with learning disabilities and students who are achieving normally. Consistently, research has shown that students without learning disabilities spell more words correctly than students with

learning disabilities (Barenbaum, Newcomer, & Nodine, 1987; Houck & Billingsley, 1989; Moran, 1981). Thus, the number of words spelled correctly has become a frequently used index of writing ability.

When using the scoring metric of total words spelled correctly, Shinn (1989) recommended counting the words that are spelled correctly and are able to stand alone in the English language. The support for using the number of words spelled correctly has followed a similar pattern as total words written. Initially, the number of words spelled correctly was viewed as a simple, yet valuable index that educators could use to measure and evaluate written expression. For example, in the Deno, Marston, and Mirkin (1982) study described earlier, moderate to strong correlations between words spelled correctly and criterion writing measures (.57 to .80) were reported. Additionally, the number of words spelled correctly discriminated between students in special education and regular education. This last finding is consistent with Tindal and Parker (1991) who also found that total words spelled correctly significantly differentiated among 260 elementary students in special education, Chapter 1, low general education, and general education. Tindal and Parker (1991) also reported a significant increase in the number of words spelled correctly across time for all groups. However, weak correlations were reported among total words spelled correctly and the language and reading portions of the Stanford Achievement Test (Gardner, Rudman, Karlsen, & Merwin, 1982) for the 80 students sampled from Grade 5 (.22 to .30) (Tindal & Parker, 1991). Criterion-validity coefficients were not reported for Grades 3 and 4.

More recently, studies have documented that total words spelled correctly was not a strong predictor of general writing performance for elementary and secondary students

(Espin et al., 1999; Espin et al., 2000; Gansle et al., 2002; Malecki & Jewell, 2003). For example, Gansle and colleagues (2002) examined several scoring indices among 83 third graders and 96 fourth graders. Gansle et al. reported weak correlations among total words spelled correctly and teacher rankings of student writing skills, scores from the Iowa Test of Basic Skills (Hoover, Hieronymus, Frisbie, & Dunbar, 1996) for the third graders, and scores from the Louisiana Educational Assessment Program (Mitzel & Borden, 2000) for the fourth grade students (.18 to .29). Thus, total words spelled correctly may not be the strongest countable index related to overall writing, regardless of age or grade.

The percentage of correctly spelled words is another index that researchers have studied. Tindal and Parker (1989) investigated whether the percentage of correctly spelled words would be a valid indicator of secondary students' writing performance. Tindal and Parker (1989) collected 6-minute writing samples from 172 students in Grades 6 through 8 who were in remedial and special educational classes. In addition to objectively scoring the writing samples, four trained judges holistically rated the writing samples on a scale from 1 (*very poor*) to 7 (*very effective*). Results from a stepwise regression indicated that the percentage of words spelled correctly was a good predictor of holistic ratings, with a correlation of .73. Additionally, the percentage of words spelled correctly discriminated between students in special and remedial classes.

Tindal and Parker's (1989) conclusions need to be viewed with caution due to a violation of the assumption of independence of variables that is required for many statistical procedures, including regression. Measurements on two or more variables collected from the same individual are likely not independent of each other. Therefore, the eight objective index scores that Tindal & Parker (1989) collected from the same

sample of individuals are not assumed to be independent. Violating this assumption may lead to inaccurate results. Unfortunately, little information is available about the effects of various degrees of nonindependence on statistical outputs (Afifi & Clark, 1997).

Additionally, since the indices were obtained from the same writing sample, several high intercorrelations existed among the independent variables. This is a problem that occurs in regression analyses and is known as multicollinearity. Multicollinearity occurs when there is an unacceptably high level of intercorrelation among the independent variables (Afifi & Clark, 1997; Tabachnick & Fidell, 2001). When multicollinearity exists, the standard errors of the regression coefficients become inflated, making it difficult to assess the relative importance of the independent variables using beta weights. Therefore, the effects of the independent variables cannot be separated. Although it is recommended that intercorrelations among independent variables not exceed .80, the statistical problems encountered when multicollinearity exists frequently occur at correlations at or above .90 (Tabachnick & Fidell). Tindal and Parker (1989) reported that the correlations coefficients among their eight objective indices ranged from .07 to .96, with five correlations above .90, indicating the presence of severe multicollinearity. Although Tindal and Parker (1989) mentioned the high intercorrelations among the independent variables, they did not correct for this limitation.

Another limitation of Tindal and Parker's (1989) study is related to the holistic scoring of the writing samples. Generally, writing products are holistically scored based upon one or more composition skills. In this study, the final products were holistically scored solely based on the students' ability to communicate. Thus, the holistic score

ignored other important aspects of writing, such as punctuation, capitalization, and cohesiveness.

In summary, the secondary school studies showed more empirical support for the use of the percentage of correctly spelled words as a valid indicator of secondary students' performance than total words spelled correctly (Tindal & Parker, 1989; Watkinson & Lee, 1992). However, there were few studies that examined the percentage of correctly spelled words for elementary students. In one study, Parker et al. (1991) reported that the percentage of correctly spelled words was suitable for screening purposes for Grades 2 to 5. They examined whether different indices appropriately dispersed in the bottom 30 to 40% of their sample's distribution. Percentile graphs with standard errors of measurement bands were created to determine whether the percentage of correctly spelled words discriminated among students at different percentile levels and between grade levels. Their sample consisted of 1,917 students enrolled in special education (7%), Chapter 1 (14%), and regular education (79%). Although other objective indices clustered at the lower end of the distribution, indicating that those scores did not distinguish among low performing students, the percentage of correctly spelled words showed a desirable amount of distribution at the lower end for all grade levels. However, Parker et al. noted that the percentage of words spelled correctly does not have enough support to be used as an assessment tool. Thus, they recommended that the search for a reliable and valid index of classroom-based writing be continued.

Correct word sequences. A third scoring metric that has been frequently examined in the literature is correct word sequences (CWS). CWS was defined by Videen, Deno, and Marston (1982, as cited in Espin et al., 2000) as two adjacent correctly

spelled words that make sense together, given the context of the sentence. Several things are taken into account when scoring CWS, including proper sentence-ending punctuation, proper punctuation in the middle of a sentence, correct capitalization of the first word in a sentence and proper nouns, and correct spelling. Researchers have theorized that CWS would be a useful indicator of written expression given that it simultaneously takes into account the number of words written and the grammar, spelling, punctuation, and capitalization of the written product (Good & Jefferson, 1998).

Videen and colleagues (1982, as cited in Espin et al., 2000) were the first to examine if CWS was a valid indicator of writing skills. Fifty participants in Grades 3 to 6 were included in the study. CWS was highly correlated with teachers' holistic ratings (r = .85) and moderately correlated with the TOWL (r = .69) and the Developmental Sentence Scoring System (r = .49). Additionally, the mean number of correct word sequences increased by grade level. Videen et al. concluded that CWS was representative of their sample's performance in written expression.

In another study examining CWS in a sample of 260 students from Grades 3 to 5, Tindal and Parker (1991) reported low to moderate validity correlation coefficients between CWS and three qualitative scores: story idea, organization, and conventions (r = .27 to .63). The three subjective measures were based on an analytic scoring system. Additionally, scores from the Stanford Achievement Test for 80 fifth grade students were correlated with CWS scores. Again, correlations were low to moderate, ranging from .31 to .41. Although correlations were at best moderate, Tindal and Parker (1991) reported a significant improvement of CWS scores from fall to spring. In addition, they found that CWS significantly differentiated between students who were learning disabled and

students who were in general education. Therefore, Tindal and Parker (1991) recommended using the number of correct word sequences as an indicator of writing performance for students who frequently make mechanical errors, such as capitalization and punctuation.

At the secondary level, a different pattern of results were found. Tindal and Parker (1989) were the first to investigate whether the number of CWS would be valid for monitoring the writing skills of secondary-level students. Tindal and Parker (1989) collected 6-minute writing samples from 172 students in Grades 6 through 8. Results from a stepwise regression indicated that the number of CWS did not predict holistic ratings of students' writing. In addition, the number of CWS did not differentiate students in special education from those in remedial education. However, Tindal and Parker (1989) reported that the percentage of CWS was a good predictor of holistic ratings (r = .75) and also that it discriminated between students in special and remedial programs. As a result, Tindal and Parker (1989) concluded that the percentage of CWS was a more valid indicator of writing performance of secondary students than the number of CWS. However, due to the previously discussed limitations of Tindal and Parker's (1989) study, results need to be viewed with caution. The high intercorrelations among independent variables and the scoring of the criterion variable may explain why a different pattern of results emerged for the number of CWS among elementary and secondary students.

In a more recent study, Weissenburger and Espin (2005) found decreasing validity coefficients for CWS as grade levels increased. In a sample of 484 fourth-, eighth-, and tenth-grade students, Weissenburger and Espin found moderate correlations between CWS and the language arts portion of a statewide assessment of achievement in fourth-

and eighth-grade students (r = .47 to .62). However, in the tenth-grade sample the validity coefficients were lower ranging from .18 to .26. Weissenburger and Espin also correlated CWS with holistic ratings of the direct writing portion of the statewide assessment. Only the fourth- and eighth-grade students completed the direct writing assessment. All validity coefficients were significant and ranged from .49 to .60. The authors of the study suggested that CWS would have been more strongly related to the holistic writing scores than the language arts portion of the statewide assessment; however, this can not be determined due to the lack of data collected for tenth graders.

In addition to the number and percentage of CWS, the number of correct minus incorrect word sequences (CWS-IWS) has been examined in the literature. This relatively new objective index was first reviewed by Espin and colleagues (2000). One-hundred and twelve students in Grades 7 and 8 completed two story writing and two descriptive writing samples. For each sample, students wrote for 3 and 5 minutes. Separate interrater reliability coefficients were calculated for story and descriptive writing samples and for 3- and 5-minute writing samples. The interrater agreement for CWS-IWS scores ranged from 88% to 92% for types and lengths of writing samples. Alternate-form reliability estimates were computed separately for story and descriptive writing samples and for 3- and 5-minute writing samples. Espin et al. reported the alternate-form reliability coefficients to be in the moderate range for type and length of writing, ranging from .72 to .78. Additional analyses were conducted to examine validity coefficients between CWS-IWS and teachers' rating of writing proficiency and a district writing test. Resulting correlation coefficients ranged from .65 to .75. Espin et al. concluded that CWS-IWS might serve as the best indicator of students' writing performance when

compared to other writing indices (e.g., total words written, number of CWS, etc.). However, it is important to note that in this study the students did not handwrite their writing samples. The students composed their writing samples on a computer system and were able to use editing features as they typed. Thus, it is possible that the use of computers may have affected the students' writing, and subsequently influenced the results of the study.

In a more recent study investigating the validity of CWS-IWS, Malecki and Jewell (2003) collected writing samples from 946 students in Grades 1 through 8. Unlike the Espin et al. (2000) study in which the writing samples were collected at one point in time, Malecki and Jewell collected writing samples once in the fall and again in the spring. A repeated measures multivariate analysis of variance was conducted and Malecki and Jewell reported that students' CWS-IWS scores were significantly higher from fall to spring for all grade levels. No grade-level interactions were found, which indicated that the changes in CWS-IWS from fall to spring were similar across grade levels. However, when each individual grade was examined, the number of CWS-IWS did not increase between fifth and sixth grade.

Espin, De La Paz, Scierka and Roelofs (2005) also examined the sensitivity of CWS-IWS to change in performance over time. Twenty-two students in the seventh and eighth grades completed an expository essay from a bank of topics. Each student was given 35 minutes to complete an essay and wrote a minimum of six essays in the beginning of the study. Following completion of the pretest essays, students were instructed in writing using composition strategies designed to help them plan, organize, and write expository essays. At the end of the 4-week instruction, students were again

asked to write expository essays. Espin and colleagues chose a random sample of three pretests and posttests from each student for data analyses. Each essay was scored for CWS and CWS-IWS. In addition, each essay was scored for the number of functional essay elements and quality ratings. Functional essay elements were defined as units in the essay that directly supported the development of the writer's paper, including premises, reasons, elaborations, and conclusions. Essays were divided into units and scored by the second author as functional or nonfunctional. Twenty-five percent of the essays were scored by an independent rater resulting in 84% scoring agreement. The overall quality of the essays was assessed using a holistic rating system of 0 (*low*) to 7 (*outstanding*). Prior to scoring, essays were typed and corrected for spelling, punctuation, and capitalization. Raters were two general education teachers who were unfamiliar with the design of the study. Raters were instructed to consider the ideas and development of the essay, the organization, unity and coherence, and the breadth of the vocabulary when scoring an essay. The interrater correlation coefficient between the two raters was .90.

Espin and colleagues (2005) conducted a multivariate analysis of variance with time as a within-subjects factor. Significant differences between pre- and posttest were found for CWS and CWS-IWS indicating that the CBM indices were sensitive to change over time. In addition, correlations between the CBM scoring indices (CWS and CWS-IWS) and the criterion measures (functional elements and quality ratings) were examined. Correlations between the CBM indices and the criterion measures were moderate to strong, ranging from .66 to .83. However, the small sample size may have affected the magnitude of the correlations.

In a larger sample of 484 fourth-, eighth-, and tenth-grade students, Weissenburger and Espin (2005) found that more complex measures, such as CWS-IWS, tend to be more strongly related to criterion measures. The researchers examined 3-, 5-, and 10-minute portions of a 10-minute writing sample. Criterion-related validity coefficients were collected between CWS-IWS and a statewide assessment of achievement. Weissenburger and Espin reported moderate correlation coefficients between CWS-IWS and the language arts subtests of the statewide assessment (r = .60 to .68) for fourth- and eighth-grade students. However, for tenth grade students, the correlation coefficients were low, ranging from .29 to .36. In addition to the above criterion-related validity coefficients, Weissenburger and Espin correlated CWS-IWS with holistic scores collected from the direct writing assessment portion of the statewide assessment. Only the fourth and eighth graders completed the direct writing assessment. All criterion-related validity coefficients were significant ($p < .001$) and ranged from .56 to .65, supporting the use of CWS-IWS as a valid indicator of writing performance for fourth and eighth grade students.

In summary, research has supported the use of the number of CWS as an indicator of writing proficiency for elementary students, but not for secondary students. However, the percentage of CWS has consistently served as a moderate to good indicator of writing performance for secondary students and has discriminated between different groups of students. Additionally, the number of CWS-IWS has shown to be a reliable and valid indicator of writing performance for both elementary and secondary students. Because these indices are gaining popularity, further research with different samples and grade

levels is needed to determine if the number and percentage of CWS and the number of CWS-IWS are reliable and valid indicators of students' writing skills.

Number of sentences. More recently, the number and types of sentences have been examined in the writing literature. Teachers theorized that the number of complete sentences may be an indicator of students' writing skills. Teachers also suggested that the types of sentences written (e.g., simple, compound, complex, and compound-complex) might be related to students' written expression skills (Gansle et al., 2002). The first study that examined the number of sentences as an indicator of writing proficiency was conducted by Espin et al. (1999). One-hundred and forty-seven students in the 10th grade completed CBM writing samples. Espin et al. defined the number of sentences as any group of words separated from another groups of words by a space, period, question mark, or exclamation point. The number of sentences was scored via a computer program. Criterion variables consisted of several measures representing the students' general writing ability. These included scores from the Language subtest of the California Achievement Test (CAT; CTB/McGraw-Hill, 1985), students' first and second semester English grades, independent ratings of the students' writing, and student group placements (e.g., an English class with students with learning disabilities, basic English, regular English, and enriched English). Students enrolled in a master of education program were the independent raters in this study. Only 45% of the writing samples were randomly selected from each of the basic, regular, and enriched groups for ratings. All of the writing samples from the group of students with learning disabilities were rated. The chosen writing samples were rated on a scale of 1 (*low quality*) to 5 (*high quality*).

The first set of analyses conducted by Espin et al. (1999) focused on the relationship between number of sentences and the criterion measures. Of all the writing indices scored (e.g., total words, CWS, characters per word, etc.) number of sentences had the highest correlations with first and second semester English grades and independent ratings of the students' writings (r = .43 to .63). When number of sentences was correlated with subtest scores of the CAT, the resulting coefficients ranged from .30 to .40. It's important to note that the CBM scores were collected a year before the CAT scores. This may have reduced the correlation coefficient between these two scores. In the second set of analyses, Espin et al. conducted a multivariate analysis of variance and reported that all four groups significantly differed on the writing indices. The univariate comparisons revealed that the four groups of students significantly differed ($p < .001$) on the number of sentences written. However, the group of students with learning disabilities only consisted of 9 students, whereas the basic English class, regular English class, and enriched English class consisted of 39, 50, and 49 students, respectively. Thus, the small number of students with learning disabilities decreases the generalizability of the results. Furthermore, no reliability data were collected for the number of sentences as an indicator of writing performance.

In a follow-up study, Espin et al. (2000) collected writing samples from 112 students in Grades 7 and 8. Each student composed two story and two descriptive writing essays. Additionally, students' written products were marked at 3 minutes and 5 minutes. Alternate form reliability was computed separately for type and length of written product. Reliability coefficients for number of sentences (i.e., any group of words separated from another groups of words by a space, period, question mark, or exclamation point) ranged

from .61 for the 3-minute descriptive writing to .82 for the 5-minute story writing. Similar results were reported by Gansle et al. (2002) who calculated an alternate form reliability coefficient for a related index, the number of complete sentences (r = .62), for students in the third and fourth grade. However, in Gansle and colleague's study the number of complete sentences only included those that started with a capital letter, had a recognizable subject, had a recognizable verb, and ended with correct punctuation. Thus, these indices are not directly comparable.

Additionally, validity coefficients were computed by Espin et al. (2000) and Gansle et al. (2002). Espin et al. calculated correlation coefficients between number of sentences and teacher ratings of students' writing and a district-wide test. Only students in Grade 8 completed the district-wide test, which consisted of students writing a response to a given scenario. The writing samples from the district-wide test were scored on a 1 (*poor writing*) to 4 (*excellent writing*) by trained raters. Although all correlations were significant ($p < .001$), they were moderate at best, ranging from .54 for 3-minute descriptive writing to .64 for 5-minute descriptive writing. Lower validity coefficients were computed between number of complete sentences and the Iowa Test of Basic Skills for third grade students, the Louisiana Educational Assessment Program for fourth grade students, and teacher rankings of students' writing skill in Gansle et al.'s study (r = .22 to .33). The differences among coefficients in both studies may have been due to the differences among the definitions for the index (i.e., the number of sentences versus the number of complete sentences).

In a more recent study, Gansle, VanDerHeyden, Noell, Resetar, and Williams (2006) examined the interrater reliability, test-restest reliability, and the criterion validity

of the number of complete sentences. Students in grades one to five ($N = 206$) completed writing samples. One hundred forty-five of these were rescored for interrater reliability. The writing samples were scored by two advanced doctoral students who were trained by the principal investigator. Interrater agreement for the number of complete sentences was 83.6%. Of the 206 students, 190 completed a second writing sample for test-retest evaluation. Test-retest reliability was computed using a Pearson correlation. The reliability coefficient was .65 between the first and second administration. However, it's important to note that the time between the first and second administration was only 1 week. Finally, in order to examine the criterion validity Gansle and colleagues (2006) correlated the number of complete sentences with scores from the Stanford Achievement Test, Ninth Edition (Stanford-9; Harcourt Brace Educational Measurement, 1996). Only 163 of the students in the second through fifth grades completed the Stanford-9. The Stanford-9 is a standardized, norm-referenced test that measures language expression, punctuation, capitalization, and grammatical concepts. The Total Language Score of the Stanford-9 incorporates all subtests measuring prewriting, composing, and editing. Although the validity coefficient was statistically significant, the relationship between the number of complete sentences and the Total Language Score was weak ($r = .36$). The low validity coefficient is comparable with results from Gansle et al. (2002) study in which they also found low validity coefficients between the number of complete sentences and standardized tests among third- and fourth-grade students.

In summary, the four studies that examined the number of sentences or complete sentences as an index of writing performance reported moderate to good alternate form reliability coefficients. However, the same studies reported low to moderate validity

coefficients between number of sentences and writing performance as measured by criterion variables such as district tests, standardized tests, and teachers' rankings. Since the number of sentences is theorized only to reflect the amount of content included in the writing product as well as the use of end punctuation to separate thoughts, more research is needed to determine the reliability and validity of sentences as an overall index of writing skill.

Correct capitalization. One of the two major mechanical skills needed for written expression is proper capitalization of words. Although proper capitalization may not enhance the meaning of a written essay, students must learn the rules of capitalization. If many capitalization errors are made, the writing may be unclear and subsequently reduce the quality of the written product (Bradley-Johnson & Lesiak, 1989; Idol, Nevin, & Paolucci-Whitcomb, 1999).

Problems with the rules of capitalization are particularly evident in students with learning disabilities. Poplin, Gray, Larsen, Banikowski, and Mehring (1980) reported significant differences in capitalization errors between students with learning disabilities and students without learning disabilities. One hundred ninety-eight students (99 students with learning disabilities and 99 normally achieving students) in Grades 3 through 8 completed the TOWL and results indicated significant differences in capitalization errors between the two groups at all grade levels. Similar results were found in a comparison study of 48 students with learning disabilities and 48 normally achieving students in Grades 4, 8, and 11. Houck and Billingsley (1989) reported that normally achieving students in Grade 4 demonstrated a 92% accuracy rate for capitalization while students in Grades 8 and 11 demonstrated a 93% accuracy rate. In contrast, students with learning

disabilities in Grade 4, 8, and 11 demonstrated 74%, 85%, and 87% accuracy rates, respectively. These results indicated that eleventh graders with learning disabilities had not achieved the accuracy that was attained by normally achieving fourth graders.

Although the percentage of correct capitalization has not been examined as a CBM index, the number of correct capitalization has been researched as an objective index of writing performance. In the Gansle et al. (2002) study, 179 third- and fourth-grade students completed 3-minute CBM writing probes on two consecutive days. Correct capitalization included words used at the beginning of a sentence and proper nouns. An alternate form reliability estimate for the correct capitalization scores for the two days was computed and was modest (r = .43). Additionally, Pearson correlations were calculated between the correct capitalization scores and teachers' rankings of student writing skill, the Iowa Tests of Basic Skills for Grade 3, and the Louisiana Educational Assessment Program for Grade 4. Correlation coefficients ranged from .15 to .26, indicating that correct capitalization was not a valid indicator of written expression as measured by the criterion variables among this sample of students.

More recently, Gansle and colleagues (2006) examined the interrater reliability, test-restest reliability, and the criterion validity of the number of correct capitalizations. Students in grades one to five (n = 206) completed writing samples. One hundred forty-five of these were rescored for interrater reliability. The writing samples were scored by two advanced doctoral students who were trained by the principal investigator. Interrater agreement for correct capitalization was 94.2%. Of the 206 students, 190 completed a second writing sample one week later for test-retest evaluation. Test-retest reliability was computed using a Pearson correlation. The reliability coefficient was low (r = .44)

between the first and second administration. Finally, in order to examine the criterion validity Gansle and colleagues correlated correct capitalization with the Total Language Score from the Stanford-9. Only 163 of the students in the second through fifth grades completed the Stanford-9. Although the validity coefficient was statistically significant, the relationship between correct capitalization and the Total Language Score was weak (r = .28). Although the interrater reliability for correct capitalization was high, the test-retest and criterion validity coefficients were low. Gansle and colleagues reported that correct capitalizations may be particularly sensitive to the types of themes students choose to write about. For example, one student may choose to write about a story referring to several individuals using their proper names; while another student may choose to write about an event that does not provide opportunities for correct capitalizations.

Correct punctuation. The second important mechanical skill needed for written expression is proper punctuation. Similar to capitalization errors, punctuation errors may reduce the quality of the final product. Punctuation rules are often arbitrary, as there is no single correct way to punctuate a given sentence. Additionally, not all punctuation is mandatory. However, simple sentences need only one ending punctuation mark while more complicated sentences require the use of commas, semicolons, and colons for independent and dependent clauses. Thus, it could be theorized that as students' writing competence increases so will the use of more complicated sentences, and as a result the number of necessary punctuation marks (Bradley-Johnson & Lesiak, 1989; Gansle et al., 2003).

Similar to capitalization problems, punctuation problems are evident in students with learning problems. Poplin et al. (1980) examined the writing performance of 198

students in Grades 3 through 8. They reported significant differences in the correct use of punctuation as measured by the TOWL between students with learning disabilities and students without learning disabilities for all grade levels. Additionally, research has reported that students who display problems with writing are less likely than students who are skilled in writing to correct their punctuation during revision. As a result, these students exhibit more punctuation errors than students who are skilled in writing (Graham & Harris, 1997; Hooper et al., 1994).

Gansle and colleagues (2002) examined correct punctuation as an indicator of writing performance. One-hundred seventy-nine students from the third and fourth grades completed 3-minute CBM writing probes on two consecutive days. Gansle et al. scored two indices related to punctuation: total number of punctuation marks and correct punctuation marks. The total number of punctuation marks was the number of punctuation marks used in the writing probe regardless of whether or not they were correctly applied. Correct punctuation marks was the total of number of punctuation marks that were in the correct location and appropriate for the particular sentence. The mean interrater agreement for total punctuation and correct punctuation marks were 91% and 86%, respectively. The alternate-form reliability coefficient for total punctuation marks was lower (r = .29) than the reliability coefficient for correct punctuation marks (r = .59). Validity coefficients were also calculated between the punctuation indices and teacher rankings of students writing skills, the Iowa Test of Basic Skills (ITBS) for third graders, and the Louisiana Educational Assessment Program (LEAP) for fourth graders. For the correlations between total punctuation marks and the criterion measures the validity coefficients were low, ranging from .18 to .32. Similar to total punctuation

marks, coefficients for correct punctuation marks and the criterion measures were low, ranging from .26 to .37.

In a more recent study, Gansle and colleagues (2006) continued to find low criterion validity coefficients for correct punctuation. In a sample of 163 second through fifth grade students, correct punctuation was correlated with the Total Language Score from the Stanford-9. The resulting validity coefficient was low (r = .39). Although the above analyses do not support the use of punctuation marks as a reliable or valid indicator of writing performance as measured by the criterion variables, Gansle and colleagues (2002) concluded that the number of correct punctuation marks may be a usable index of writing skill. They based their conclusion on results from a series of stepwise regressions that was conducted for the ITBS scores from the Language Usage/Expression subscale and the ITBS Language Total score for third graders. For students in Grade 4, a series of stepwise regressions were conducted for the LEAP scores from the Write Competently subscale and the Use of Conventions of Language subscale. The predictor variables were 17 writing indices, 12 hand-scored writing indices (e.g., total words written, words spelled correctly, number of correct punctuation, etc.) and 5 computer-scored writing indices (e.g., MS Word Flesch Reading Ease, WP sentence complexity, etc.). However, similar to other CBM studies, multicollinearity was likely a problem in this study. Although Gansle et al. did not report intercorrelations among their independent variables, past research has documented high correlations among the writing indices (Malecki & Jewell, 2003; Tindal & Parker, 1989; Watkinson & Lee, 1992). Additionally, since the indices were all obtained from the same writing sample, it is more than likely that many of the indices would be highly correlated with each other. As a

result, it is possible that the multicollinearity confounded the results reported by Gansle et al., thereby decreasing confidence in their conclusions.

Research on combined writing indices. It has been theorized that a combination of writing indices may predict a writing criterion better than any single index (Espin et al., 1999). However, if educators are to take the time to score students' writing compositions using different indices, they need to be certain that the writing indices used are independent of each other. Using a combination of writing indices to score a written product would be useless if all the indices are contributing the same information. Therefore, each index that is measured should add new and valuable information to the assessment.

Several studies have established strong relationships among the different objective indices used in CBM in writing (Jewell & Malecki, 2005; Malecki & Jewell, 2003; Tindal & Parker, 1989; Watkinson & Lee, 1992). Using factor analytic techniques to help group the indices that are highly intercorrelated will aid in our understanding of what the specific indices are contributing to the overall assessment. Only two studies to date have factor analyzed the different writing indices. Tindal and Parker (1989) conducted a principal axis factor analysis with varimax rotation on eight of the scoring metrics: total words written, number of words spelled correctly, CWS, legible words, mean length of CWS, percentage of words spelled correctly, percentage of CWS, and percentage of legible words. Participants included 172 students (30 students from special education and 142 students from remedial programs) in Grades 6 through 8. Tindal and Parker (1989) reported that two factors accounted for 83% of the total variance: production dependent (total word written, legible words, words spelled correctly, CWS)

and production independent (mean length of CWS, percentage of words spelled correctly, percentage of CWS, percentage of legible words) metrics. The authors explained that production dependent metrics are dependent on the amount of writing and are more closely associated with fluency in writing. In contrast, the production independent scoring metrics are not dependent on the length of the final product, but are more closely related to accuracy in written expression (Tindal & Parker, 1989; Watkinson & Lee, 1992). Unfortunately, the results reported in Tindal and Parker's study (1989) may not be generalizable to students in regular education classes, as only students with learning problems were included in the study.

A second factor analytic study was conducted by Tindal and Parker (1991) using a sample of 260 students from Grades 3 to 5. A principal components analysis with varimax rotation was conducted on nine writing indices (six CBM writing indices and three analytical scoring indices). Tindal and Parker (1991) reported a rather clear and simple structure, with three factors accounting for 81% of the total variance. The first factor, accounting for 37% of the variance, included four production dependent variables: total words written, correctly spelled words, CWS, and the number of CWS-IWS. The second factor, accounting for 26% of the variance, consisted of only two variables: percentage of CWS and subjective judgment of writing conventions. Finally, the third factor accounted for 18% of the variance and was defined by the other two subjective scores: story idea and organization-cohesion. The number of incorrect word sequences did not saliently load on any of the factors.

The first two factors of Tindal and Parker's (1991) study were similar to the earlier factor analytic study conducted by Tindal and Parker (1989). In both studies, the

first factor included indices related to fluency, such as total words written, and the second factor consisted of indices related to accuracy, for example percentage of words in correct sequence. Although results are similar in both studies, they need to be examined with caution. Researchers conducting factor analyses should carefully consider many details, such as the conditions that may affect factor analytic results (e.g., multivariate normality, sampling adequacy, multicollinearity, etc.). Other crucial decisions that need to be made a priori include the choice of a factor model, extraction methods, and rotation methods. Additionally, researchers need to decide which procedures to use to determine the optimal number of factors to retain and they need to decide a priori how they plan to interpret the factors that they obtain. It is the researchers' responsibilities to clearly communicate the decisions made with enough detail to allow an informed review (Gorsuch, 1997). Unfortunately, Tindal and Parker (1989, 1991) did not adequately describe their methods. For example, in both studies Tindal and Parker (1989, 1991) did not report how they determined the number of factors to retain.

Another limitation of the studies (Tindal & Parker, 1989, 1991) was the small sample sizes. For factor analytical investigations, sample sizes smaller than 300 may be inadequate (Comrey & Lee, 1992; Gorsuch, 1997). Additionally, both factor analyses were conducted using varimax rotation. Varimax is a type of orthogonal rotation, which assumes the factors to be uncorrelated (Tabachnick & Fidell, 2001). However, the CBM writing indices are expected to be correlated with each other, and in fact, Tindal and Parker (1989) discussed the high intercorrelations among the variables. An oblique rotation, which assumes the factors are correlated, might have provided a more accurate picture of the solution.

Tindal and Parker (1989) provided the correlation matrix for their data in their first factor analytic study. Due to the limitations of their study, the data were independently reanalyzed using principal components analysis with oblimin rotation. A simple structure that was very similar to Tindal and Parker's (1989) factor solution was obtained, with two components accounting for 86% of the total variance. The first component accounted for 57% of the total variance and consisted of total words written, legible words, words spelled correctly, and CWS. The second component accounted for 29% of the total variance and consisted of mean length of CWS, percentage of words spelled correctly, percentage of CWS, and percentage of legible words. All indices saliently loaded on a single component and all loadings were above .70. The similarity between the factor analytic results supports the hypothesis that one set of indices is measuring a writing fluency component, while another set of indices is measuring a writing accuracy component. Unfortunately, the correlation matrix for the data of Tindal and Parker's (1991) second study was not provided. Thus, further post-hoc analyses were not possible.

A more recent study examined how production dependent and production independent indices relate to criterion measures (Jewell & Malecki, 2005). Two hundred and three students in Grades 2, 4, and 6 completed one 3-minute CBM writing probe. The CBM writing probes were scored for production dependent indices and production independent indices. Jewell and Malecki calculated the production dependent composite by taking the average of students' raw scores on total words written, words spelled correctly, and correct word sequences. They calculated the production independent composite by taking the average of students' scores on the measures of percentage of

words spelled correctly and percentage of correct writing sequences. In addition, the writing samples were scored using the Tindal and Hasbrouck Analytic Scoring System (THASS; Tindal and Hasbrouck, 1991). Using this system, students' writing samples were scored on a 5-point scale on three dimensions of writing: story-idea, organization-cohesion, and conventions-mechanics. Students were also administered the Stanford Achievement Test (SAT; Harcourt Brace Educational Measurement, 1996), which is a standardized, norm-reference test that measures Language and Spelling, among other academic areas. Finally, students' fall semester grades (the average of quarter 1 and quarter 2 grades) for Language Arts were also examined for fourth and sixth grade students. Students' grades included plusses and minuses; however, any "A" was converted to a 4.0, any "B" to a 3.0, etc.

Correlational analyses between students' scores on the CBM composites, total THASS scores, subtest scores on the SAT, and Language Arts grades were conducted. The validity coefficients among the CBM composites and the THASS scores were significant for students in Grades 2 and 4 (r = .41 to .51). However, for Grades 6 the validity coefficient between the production dependent composite and the THASS scores was not significant (r = .31) while the validity coefficient between the production independent composite and the THASS score was significant (r = .50). When comparing the CBM composites to an analytic scoring system, the relationship was moderate at best.

As in the above analyses, the correlation coefficients among the CBM composites and the language subtest scores on the SAT were significant for Grades 2 and 4. However, the coefficients between the production dependent composite and the language subtests scores were low (r = .34 to .42) compared to moderate coefficients between the

production independent composite and the language subtest scores (r = .58 to .65). The validity coefficient between the production dependent composite and the language subtest score for sixth grade students was not significant (r = .03) while the validity coefficient between the production independent composite and the language subtest score for sixth grade students was significant (r = .54). A similar trend was seen for the CBM composites and spelling subtest scores on the SAT. The validity coefficients between the production dependent composites and the spelling subtests for all three grade levels were not significant (r = .17 to .20). However, the validity coefficients between the production independent composites and the spelling subtests for all three grade levels were significant (r = .48 to .55).

Lastly, Jewell and Malecki (2005) correlated the CBM composites with Language Arts Grades for the fourth- and sixth-grade students. For fourth-grade students the validity coefficients were significant for production dependent composite (r = .53) and for production independent composite (r = .60). However, for the sixth-grade students the validity coefficients between the CBM composites and Language Arts Grades were not significant and ranged from .22 to .35.

Jewell and Malecki (2005) also conducted a series of four regression analyses to determine which type of curriculum-based index was most related to students' scores on the THASS. The predictor variables included students' scores on the production dependent composite, production independent composite, correct minus incorrect word sequences, and their grade level. The four predictor variables collectively accounted for 53% of variance in students' Total THASS scores ($p < .01$). Correct minus incorrect word sequences was the only predictor that significantly contributed to the Total THASS

scores (β = .48, p < .01). Regression analyses conducted with students' story-idea THASS scores revealed that the four predictor variables accounted for 35% of the variance (p < .01), with production dependent composite being a significant predictor (β = .43, p < .01). The four predictor variables collectively accounted for 30% of the variance in students' scores on organization-cohesion (p < .01), but none of the individual variables were significant predictors. Finally, a total of 62% of the variance in students' scores on conventions-mechanics was accounted for by the four predictor variables (p < .01). Correct minus incorrect word sequences (β = .77, p < .01) and students' grade level (β = .22, p < .01) were both significant predictors.

In summary, Jewell and Malecki (2005) found that measures of writing accuracy (i.e., production independent indices) may be more strongly related to students' performance on different types of writing criteria than measures of writing fluency (i.e., production dependent indices). These results are consistent with past research (Tindal & Parker, 1989) that revealed production independent scoring indices were more closely related to teachers' holistic ratings than production dependent scoring indices.

Summary of writing indices. The CBM written expression literature suggests that some written expression indices are differentially effective for elementary and secondary students. Generally, three writing indices have been examined at the elementary level: total words written, total words spelled correctly, and the number of CWS. These production dependent scoring indices are fluency-based (Tindal & Parker, 1989, 1991; Watkinson & Lee, 1992). In the majority of research with elementary students, total words written was the strongest and most consistent index correlated with writing criteria (Deno, Marston, & Mirkin, 1982; Deno, Marston, Mirkin, Lowry, et al. 1982; Marston et

al., 1981). Initial research found total words spelled correctly to be a strong predictor of general writing performance for elementary students (Deno, Marston, & Mirkin; Tindal & Parker, 1991), but more recent studies with elementary students reported that total words spelled correctly did not predict criterion measures of writing (Gansle et al., 2002). Thus, results for this index have been incongruent across time. Finally, the number of CWS was reported to correlate at a moderate to high level with criterion measures of writing in several studies (Espin et al., 2000; Tindal & Parker, 1991), and were shown to discriminate between students in a learning disabilities classroom and a general education classroom (Tindal & Parker, 1991). Additionally, the number of CWS were reported to increase from Grades 3 through 6 (Tindal & Parker, 1991) and from Grades 1 through 5 (Malecki & Jewell, 2003). Thus, research at the elementary level has favored total words written and the number of CWS as reliable and valid indicators of students' writing performance.

At the secondary level, production dependent indices have consistently failed to reliably or validly measure students' performance in written expression (Espin et al., 2000; Espin et al., 1999; Gansle et al., 2002; Malecki & Jewell, 2003; Parker, Tindal, & Hasbrouck, 1991; Tindal & Parker, 1989; Watkinson & Lee, 1992; Weissenburger & Espin, 2005). In contrast, research has supported the reliability and validity of production independent indices, or measures of accuracy, as indicators of students' writing performance (Espin et al., 2005; Parker, Tindal, & Hasbrouck; Tindal & Parker, 1989; Watkinson & Lee). In particular, the percentage of correctly spelled words and the percentage of CWS have moderately predicted writing criteria. The same indices significantly discriminated between students with and without learning disabilities and

significantly discriminated between students in special and remedial education (Tindal & Parker, 1989; Watkinson & Lee). In addition to production independent indices, a relatively new index, CWS-IWS, has been a moderate to good indicator of writing performance for secondary students (Malecki & Jewell; Espin et al., 2000; Espin et al., 2005). Preliminary evidence has also shown the number of CWS-IWS to significantly increase from fall to spring in Grades 1 through 8 (Malecki & Jewell), demonstrating sensitivity to change over time. Thus, at the secondary level the percentage of correctly spelled words, the percentage of CWS, and the number of CWS-IWS are the most reliable and valid indicators of students' writing performance.

PURPOSE OF PRESENT STUDY

Initially, CBM was developed so that special education teachers could monitor elementary students' growth in basic skill areas, including reading, math, and writing. Thus, most of the CBM research was conducted at the elementary school level. As CBM developed through the years and became more popular, teachers have become interested in applying similar formative procedures with students at the secondary school level. Therefore, technical work has begun on identifying CBM writing indices that are reliable and valid for assessing and monitoring student growth with older students (Deno, 2003; Espin et al., 2000).

In general, the psychometric properties of CBM written expression indices are weaker than 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. Additionally, identification of a writing criterion that is widely accepted as a good measure of writing is controversial (Cole, Haley, et al., 1997; Espin et al., 2000; Gansle et al., 2002; Hooper et al., 1994). However, some writing indices have shown potential for screening and progress monitoring as evident by the reliability and validity evidence. Thus, further research is needed to determine if there are one or more writing indices that can be used as an effective and feasible measure of writing for secondary students (Espin et al., 1999).

Consequently, the purpose of the present study was to investigate 10 CBM measures of written expression for a sample of students in Grade 8. Although there are several measures available for assessing writing skills, many are subjective, difficult to score, and time-consuming (Watkinson & Lee, 1992). If findings support the use of a

curriculum-based scoring index, school personnel can utilize these scores as a valid and efficient tool for assessing and monitoring progress of students' writing skills. Past research has suggested that specific CBM indices may be more suitable for certain age groups. As age increases, the students' writing becomes more complex. Thus, research is needed to examine the validity of CBM writing indices with respect to higher levels of complex writing. Additionally, past research most often included criterion measures which represented either direct or indirect assessments of writing, but not both forms of writing. Thus, this study utilized a criterion measure that has both forms of writing. Given this purpose, the following research question was posed: Which CBM index best predicts an accepted criterion measure of writing for eighth-grade students?

METHOD

Setting

The study was conducted in a metropolitan city in central New Jersey. According to the 2000 census, the population of the city is estimated to be about 40,000 people and the median family income is $66,266. The racial/ethnic composition includes 76% White, 8% Black, 1% American Indian/Alaska Native, 7% Hispanic/Latino, 10% Asian, less than 1% Pacific Islander, 2% multiracial, and 2% other. The above percentages do not equal 100% due to Hispanic and Latinos belonging to other races. Of the 7% Hispanic/Latino population, 5% are White, less than 1% are Black, less than 1% are American Indian/Alaska Native, and 2% are of other races. Approximately 77% of the population speaks English as their first language (U.S. Census Bureau, 2000).

The present school district operates four elementary schools (Kindergarten through 3rd grade), an upper elementary school (4th and 5th grades), a middle school (6th through 8th grades), and a high school (9th through 12th grades) with a total enrollment of about 5,000 students. In the 2004 – 2005 school year, 16% of the school's population was in special education. Approximately 76% of the school population's primary language was English and 2% of the student body was classified as a Limited English Proficient (LEP) learner. In regards to the socioeconomic status of the students attending this school district, 9% of the students received free lunch and 15% of the students received reduced lunch. (New Jersey Department of Education, 2003).

Participants

Participants included 447 eighth-grade students. Participants in this sample ranged from ages 12 to 16, with 66% of the sample aged 13 and 30% of the sample aged 14. The

sample was 60% Caucasian, 14% Black, 14% Hispanic, and 12% Asian/Pacific Islander. Fifty-two percent of the sample was male and 48% was female. Sixty-six students, 15% of the sample, were classified as special education and 381 students, 85% of the sample, were in regular education.

Procedure

After obtaining consent from the school's superintendent and board of education, all eighth-grade students were administered the TOWL-3 and CBM writing probe as practice for the state-wide assessment test. There were 463 students enrolled in the eighth grade. Sixteen of the eighth-grade students did not participate in the study due to absence, suspension, or lack of English language skills. One student did not participate because of a broken hand. The remainder, 447 students, completed the TOWL-3 and CBM writing probe over two or three 40-minute class periods.

Measures

Predictor Variables

The predictor variables in this study were the various writing indices used for scoring CBM writing probes, including: total words written, words spelled correctly, percentage of words spelled correctly, number of CWS, percentage of CWS, number of CWS-IWS, number of sentences, number of correct capitalization, number of punctuation marks, and correct punctuation marks. Detailed descriptions, examples, and scoring guides for each index are included in the training manual (see Appendix).

Total Words Written

Total words written was defined as the number of words that the student wrote in 3 minutes. Spelling, grammar, and content were not taken into consideration when

counting the number of words. Numerical representations and symbols were not included in this total.

Words Spelled Correctly

Total words spelled correctly was defined as the number of correctly spelled words written by the student. Each word counted as correct had to be able to stand alone in the English language. However, context and grammar were not taken into account. Therefore, the word did not need to be used correctly within the context of the sentence, it only needed to be spelled correctly.

Percentage of Words Spelled Correctly

The percentage of words spelled correctly was the ratio of the number of words spelled correctly to the total number of words written in the composition.

Number of CWS

CWS was defined as two adjacent, correctly spelled words that were syntactically and semantically appropriate given the context of the sentence. Thus, words were examined for correct meanings, tenses, number agreement (singular or plural), and noun-verb correspondences, when identifying CWS. In addition, punctuation, capitalization, and spelling were taken into account when scoring correct word sequences.

Percentage of CWS

The percentage of CWS was the ratio of the number of correct word sequences divided by the total number of possible word sequences.

Number of CWS-IWS

The number of CWS-IWS was calculated by subtracting the number of incorrect word sequences from the total number of correct word sequences.

Number of Sentences

A sentence was defined as any series of words separated from another series of words by a space or punctuation mark, such as a period, question mark, or exclamation point. The series of words had to include a recognizable subject and verb, but did not need to contain the appropriate beginning capitalization or correct ending punctuation. The total number of sentences written by the student was the number of sentences index.

Number of Correct Capitalization

Correct uses of capital letters were counted to determine the number of correct capitalization. This included the first letter of a word used to begin a sentence, days of the week, months, holidays, countries, languages, nationalities, religions, people's names and titles, trade-marks and names of companies, places and monuments, names of vehicles, and titles of books, poems, songs, plays, and films. Additionally, capital letters were required for the personal pronoun 'I' and for acronyms.

Number of Punctuation Marks

The number of punctuation marks was defined as the total number of punctuation marks used, regardless of whether they were appropriate for the sentence.

Number of Correct Punctuation Marks

The number of correct punctuation marks was defined as only those punctuation marks that were determined to be used appropriately for that sentence. Additionally, students had to correctly place the punctuation mark in the sentence.

Criterion Variable

Although it is difficult to identify a writing criterion that is unanimously accepted as a good measure of writing (Cole, Haley, et al., 1997; Espin et al., 2000; Gansle et al.,

2002; Hooper et al., 1994), there are several assessment factors that should be considered when selecting a measure of written expression. First, in order to appropriately evaluate student performance, the data obtained from an assessment method should be reliable and valid (Cole, Muenz, Ouchi, Kaufman, & Kaufman, 1997; Salvia & Ysseldyke, 2001). Second, Hooper et al. recommended that writing assessments should include production components, such as spelling, proofreading, and mechanics in order to determine the students' knowledge of writing conventions. Third, writing assessments should contain some aspect of direct measurement, in which students can apply writing conventions to an actual writing task (Cole, Muenz, et al.; Hooper, 2002; Hooper et al.); thereby increasing the ecological validity of the assessment. Fourth, Hooper and colleagues suggested that a picture stimulus be used to elicit a writing sample. This stimulus should be a color photograph, contain at least two characters, display a novel and interesting depiction, and portray a state of conflict. Fifth, the writing measure should contain scoring criteria that differentiate among poor and skilled writing qualities (Cole, Muenz, et al.).

Given these assessment recommendations, the Test of Written Language – Third Edition (TOWL-3; Hammill & Larsen, 1996) was chosen as the criterion variable. The TOWL-3 is an individually or group administered comprehensive test of writing for children between the ages of 7 and 17 (Hammill & Larsen, 1996). 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 authors reported that the

underlying model of the TOWL-3 measures three components of writing: conventional, linguistics, and cognitive. The conventional component refers to the ability to write in compliance with accepted standards of the English language, including the efficient and consistent use of the rules for punctuation, capitalization, and spelling. The linguistic component refers to the syntactic, morphologic, and semantic elements of English, including the appropriate selection of suitable words, tenses, plurals, noun-verb correspondences, and cases. The cognitive component refers to the ability to express ideas, opinions, and thoughts in a logical and coherent manner. The TOWL-3 is comprised of eight subtests, five that incorporate an indirect assessment format and three that incorporate a direct assessment format (Hammill & Larsen).

The first five subtests, Vocabulary, Spelling, Style, Logical Sentences, and Sentence Combining require the student to demonstrate writing skills in isolation. These subtests measure discrete elements of the English language and ignore the overall quality of the written message. The Vocabulary subtest requires the students to write a sentence that incorporates a stimulus word. In the Spelling and Style subtests the students are required to write sentences from dictation. The students are told to pay particular attention to their spelling, punctuation, and capitalization. The Logical Sentences subtest asks the students to edit illogical sentences into sentences that make sense. In Sentence Combining, the students are required to combine several short sentences into one grammatically correct written sentence. There are no time limits for any of these subtests.

The three remaining subtests, Contextual Conventions, Contextual Language, and Story Construction, require the students to apply their writing knowledge to a spontaneous writing task. For these subtests, students have 15 minutes to write a story

from a picture. There are two black-and-white, stimulus pictures. Picture one contains a futuristic scene with astronauts, space ships, and construction activity. Picture two contains a scene of people dressed in animal skins hunting a prehistoric mammal. Depending on which form of the TOWL-3 is being used, students write a story about one of these pictures. The examiner evaluates the writing sample for various criteria. To earn points in the Contextual Conventions the students must demonstrate knowledge of punctuation, capitalization, and spelling rules in their writing samples. For Contextual Language, the examiner evaluates the writing sample for vocabulary, sentence construction, and grammar. For Story Construction, the students' writing samples are evaluated for the quality of the plot, writing style, development of characters, interest, and other compositional aspects.

Standard scores and percentiles are available for each individual subtest. In addition, three composite quotients may be computed. The Overall Writing Quotient is calculated by using the standard scores from all eight subtests and represents the student's overall writing proficiency. The authors of the TOWL-3 asserted that the Overall Writing Quotient is the best estimate of a student's general ability in written expression. The Contrived Writing Quotient represents performance on the first five subtests (Vocabulary, Spelling, Style, Logical Sentences, and Sentence Combining) and estimates the student's knowledge about elements of writing. The Spontaneous Writing Quotient represents performance on the last three subtests (Contextual Conventions, Contextual Language, and Story Construction) and estimates the student's ability to write a spontaneously composed essay. All three quotients have a mean of 100 and a standard deviation of 15.

Reliability

The concept of reliability refers to the consistency with which tests scores are stable over repeated applications (AERA, APA, & NCME, 1999). If test scores have good reliability, there is less error associated with their scores. Therefore, the study of reliability centers on estimating the degree of error associated with the test scores. Hammill and Larsen (1996) examined three sources of error variance that can affect the reliability of the TOWL-3 scores: content sampling, time sampling, and interscorer differences.

Content sampling. Internal consistency reliability refers to the procedure that is used to determine the amount of error due to content sampling. One type of internal consistency reliability is calculated by Cronbach's coefficient alpha. This type of internal consistency demonstrates the degree of homogeneity among test items. Alpha coefficients were calculated for each age group using the data from the normative sample. For all ages, the coefficient alphas for the scores obtained from the Overall Writing Quotient and the Contrived Writing Quotient were all above .90, exceeding the recommended value for individual-decision making (Salvia & Ysseldyke, 2001). The coefficient alpha for scores obtained from the Spontaneous Writing Quotient was also .90 or above for ages 9 through 17. For students aged 7 and 8, the coefficient alpha for the scores obtained from the Spontaneous Writing subtest were slightly lower at .89. Additionally, internal reliability coefficients were calculated for the individual subtests for all age groups. Coefficient alphas ranged from .60 for children aged 7 for Contextual Conventions to .94 for children aged 16 for Story Construction, with most correlation coefficients above .80. Coefficient alphas for selected subgroups of gender, ethnicity, and

disability status for the composite scores were not provided. However, reliability coefficients for these subgroups and the individual subtests were calculated and ranged from .65 to .95, with most reliability coefficients falling in the acceptable range of .80 (Salvia & Ysseldyke) or higher (Hammill & Larsen, 1996).

Alternate-forms reliability is another procedure that can be used to determine the amount of error due to content sampling. Both forms of the TOWL-3 were completed by students from the normative sample during one testing session. The scores from both forms were then correlated to estimate the amount of content sampling error. The authors did not provide any information regarding the sample size, time between administrations, and format of administration (individual or group). With one exception, alternate-form reliability estimates were in the acceptable range of .80 or higher for all composite scores for all ages. The one exception was for children aged 17, in which the two sets of scores for the Spontaneous Writing Quotient was .76. The average alternate-form correlation coefficient for the scores obtained from the Contrived Writing, Spontaneous Writing, and Overall Writing composites were .92, .83, and .93, respectively. Overall, the alternate-form reliability estimates were comparable to the coefficient alphas reported previously. Thus, the authors have demonstrated evidence of high internal consistency of the TOWL-3 composite scores.

Time sampling. Time sampling examines the extent to which a student's test performance is consistent over time. It is generally measured by calculating a correlation coefficient between a student's scores on a given test administered at two separate points in time (AERA, APA, & NCME, 1999). Hammill and Larsen (1996) investigated test-retest reliability in two separate groups of students. One group was comprised of 27

students in Grade 2 and the other group consisted of 28 students in Grade 12. The groups were administered both forms of the TOWL-3 two weeks apart. The mean test-retest correlation coefficients for forms A and B for children in Grade 2 were .85 for the Spontaneous Writing composite scores and .86 for the Contrived Writing and Overall Writing composite scores. The mean test-retest correlation coefficients for forms A and B for students in Grade 12 were .90, .88, and .92 for the Contrived Writing, Spontaneous Writing, and Overall Writing composite scores, respectively.

Although these coefficients are viewed as acceptable test-retest reliability estimates (Salvia & Ysseldyke, 2001), the methods used by Hammill and Larsen (1996) may be problematic. First, only 2% of the normative sample was used in the test-retest analysis. Second, only two grades, Grade 2 and 12, were examined. Third, both grades were from only one geographic area of the nation (Cole, Haley, et al., 1997).

Interscorer differences. Interscorer differences are more common in tests that involve subjective judgments than those that rely on objective judgments. The TOWL-3 has several subtests that require subjective scoring. Two staff members of the test's publishing company, PRO-ED, who were familiar with the test's scoring procedures independently rated 38 TOWL-3 protocols drawn at random from the normative sample. The results of the two scorings were correlated for each of the TOWL-3 subtests and composite quotients. The interscorer reliabilities for Form A for the subtests ranged from .80 for Story Construction to .96 for Vocabulary and Spelling. The interscorer reliabilities for Form B for the subtests ranged from .86 for Story Construction and Logical Sentences to .97 for Spelling. For the composite scores, the mean interscorer reliabilities were .92

for Spontaneous Writing and .97 for Contrived Writing and Overall Writing (Hammill & Larsen, 1996).

All the interscorer reliabilities coefficients fell in the acceptable range of .80 or higher. However, it is difficult to determine if these estimates would be obtained by typical test users as the two raters were staff members of PRO-ED. PRO-ED staff members are not the typical users of the TOWL-3. Also, the level, type, and amount of training these two raters received were not described making it difficult to generalize results of this study (Bucy & Swerdlik, 1998; Cole, Haley, et al., 1997).

Validity

Validity refers to the degree to which different types of accumulated evidence support the intended interpretation and use of the test scores (AERA, APA, & NCME, 1999). The authors of the TOWL-3 explored the content-related, criterion-related, and construct-related validity evidence of its scores (Hammill & Larsen, 1996).

Content-related evidence. 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 (AERA, APA, & NCME, 1999). It is important that the item content of TOWL-3 adequately sample the main components of written language, which the authors have identified as conventional, linguistic, and conceptual. Thus, Hammill and Larsen (1996) presented a strong and clear rationale for each subtest and the procedures used to select its items (Bucy & Swerdlik, 1998; Hansen, 1997).

Additionally, classical item analysis was used to screen items during test development. Specifically, 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 authors included only those items that had a correlation coefficient of .30 or higher and were statistically significant. In Hammill and Larsen's (1996) final item analysis study, 93% of the median discrimination correlation coefficients for students aged 9 through 17 fell in the author's acceptable range of ≥ .30. However, for students aged 7 and 8, the median discrimination correlation coefficients were unacceptably low, with only 38% at or above .30. Item difficulty, the percentage of students who passed a specific item, was also examined to identify items that were too easy or too hard for the Vocabulary, Spelling, Style, Logical Sentences, and Sentence Combining subtests. Hammill and Larsen accepted items that had a distribution between 15% and 85%. Eighty-three percent of the median percentages of difficulty for the items were in the accepted range. Again, most of the unacceptable percentages were for students aged 7 and 8. Thus, the TOWL-3 may be inappropriate for use with students aged 7 and 8, and more appropriate for students aged 9 through 17.

Content-related validity of the TOWL-3 was also assessed with Differential Item Functioning (DIF) analysis in order to detect biased items. Hammill and Larsen (1996) used the Delta Scores approach that was developed by Jensen (1980). 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. Thus, the authors demonstrated that the TOWL-3 had little or no item bias within the three groups investigated.

Criterion-related evidence. Criterion-related evidence attempts to demonstrate that test scores are related to some other criterion variable that is thought to measure a similar construct (AERA, APA, & NCME, 1999). 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 Writing Scale of 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, including writing. Low to moderate correlations were reported between the composites of the TOWL-3 and the Writing Scale of the CSSA, with correlation coefficients ranging from .46 to .55 for Form A and .50 to .59 for Form B.

In addition to Hammill and Larsen's (1996) criterion-related validity study, Burns and Symington (2003) examined the correlations between the Spontaneous Writing Quotient and teacher ratings of progress in the local curriculum for 147 third- and fifth-grade students. The students' teachers were asked to rate each student's writing skills on a scale of 1 (*significant difficulties with the writing curriculum*) to 5 (*mastered the local writing curriculum*). Teachers were instructed by the authors to consider the content of the student's writing, the language used, and the conventions of written expression. The corrected correlation coefficient between the Spontaneous Writing Quotient and teacher ratings was .47. The validity coefficients between teacher ratings and the Contextual Conventions, Contextual Language, and Story Construction subtests were .48, .39, and .40, respectively. Similar to Hammill and Larsen's findings, the results from Burns and

Symington's study offer moderate support for the criterion-related validity evidence of the TOWL-3 as a measure of writing when compared to teacher ratings.

Construct-related evidence. Construct-related evidence refers to the extent to which a test measures a theoretical construct or model (Salvia & Ysseldyke, 2001). Hammill and Larsen (1996) generated hypotheses regarding the underlying construct of the TOWL-3 and used logical or empirical evidence to try to support these hypotheses. 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, which is usually during the secondary school years. 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, leveling off after age 13. Correlation coefficients showing the relationship 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 demonstrate that the relationship among the TOWL-3 subtests is moderately high.

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 (Hammill & Larsen, 1996). 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. In a review of 12 published written expression measures, Cole, Haley, et al. (1997) indicated that the TOWL-3 was the only written language test to meet the criteria of appropriately differentiating between average and below average writing skills.

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 (Reading, Math, and General Facts) in a sample of 76 students. Correlation coefficients between the composite quotients of the TOWL-3 and the three subscales of the CSSA 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 (Hammill & Larsen, 1996). 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.

Sixth, the authors of the TOWL-3 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. Hammill and Larsen conducted a principal components analysis on the data obtained 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. Unfortunately, no other information regarding their methods was provided. Additionally, the correlation matrix was not provided, therefore no post-hoc analyses could be conducted to support their findings. Therefore, these results should be viewed with caution.

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 (Hammill & Larsen, 1996). This is also known as an item's discriminating power and has

been discussed in the previous section. Hammill and Larsen demonstrated that 83% of the resulting correlation coefficients fell in the acceptable range of .30 or higher. Therefore, they 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.

Summary of the TOWL-3

Overall, the psychometric properties of the TOWL-3 are sufficient. The internal consistencies of composite and total scores are high enough for use in making individual decisions and the interscorer reliability is quite good for this type of test (Salvia & Ysseldyke, 2001). The test's content appears to be logical and well conceived (Bucy & Swerdlik, 1998; Hansen, 1997; Salvia & Ysseldyke) and the authors have provided different types of evidence supporting the underlying construct of the test (Hammill & Larsen, 1996). Therefore, the TOWL-3 meets the first recommendation that a writing measure should have sound psychometric properties when evaluating student performance (Cole, Muenz, et al., 1997; Salvia & Ysseldyke).

The second and third recommendations were that a writing measure includes both indirect assessment (Hooper et al., 1994) and direct assessment (Cole, Muenz, et al., 1997; Hooper, 2002; Hooper et al.) of writing. Of the 12 written expression measures reviewed by Cole, Haley, et al. (1997), the TOWL-3 was one of only three tests that included both direct and indirect assessment methods. The fourth recommendation was with regard to the stimulus picture. In the review conducted by Cole, Haley, et al., they found that the stimulus pictures of the TOWL-3 met three of the four stimulus criteria recommended by Hooper et al. (1994). The one exception was that the TOWL-3 stimulus pictures are black-and-white, and not in color, as recommended. Finally, of the 12 writing

assessments reviewed by Cole, Haley, et al., the TOWL-3 was the only measure to include scoring criteria that differentiated between students with average writing skills and students with poor writing skills. Because the TOWL-3 satisfied most of the recommendations for a well designed written expression measure it was used as the criterion variable in this study.

Scoring

Ten doctoral students in school psychology were trained in a 4-hour session to score the ten CBM indices and the TOWL-3. Scorers received a combination of two training manuals (Powell-Smith & Shinn, 2004; Wright, 1992) that included detailed descriptions and scoring instructions for the ten CBM indices. Additionally, for the CWS, punctuation, and capitalization indices, the scorers received a lengthy handout that reviewed the rules of grammar (APA, 2001; Grammar Slammer, 1997). Finally, the scorers utilized the TOWL-3 Examiner's Manual (Hammill & Larsen, 1996) to learn how to score the TOWL-3 protocols. The scorers were initially trained on three sample CBM probes and two sample TOWL-3 protocols. Once trained, the scorers were tested for accuracy with a final protocol of each test. If 95% or greater accuracy was attained on this final product then the raters proceeded to score actual student protocols. However, if 95% accuracy was not attained, the raters were retrained until 95% or greater accuracy was reached.

To ensure scoring consistency throughout the scoring period, every eighth protocol from each rater was checked by the primary investigator to identify common errors. If errors were found, the packet of eight protocols was sent back to that rater for the individual to rescore. Additionally, scorers posed specific scoring questions to the

investigator via email throughout the scoring process. The emailed questions and the responses were forwarded to the other scorers to promote scoring consistency. Once all protocols were received, the principal examiner rechecked all protocols for clerical and computational errors (e.g., incorrect addition of raw scores, incorrect discontinuation rules, etc.). In addition, the principal researcher's data was checked against the independent scorers' data. There were 33 cases in which the two scorers disagreed by more than 2 points. These 33 cases were rescored to ensure accurate results. Of the 33 cases, 21 resulted in modified scores.

Data Analysis

A multiple regression analysis was completed to determine which CBM index best predicted written expression, as measured by the TOWL-3. Simply stated, multiple regression combines two or more predictor variables to predict a value on a criteria 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. These regression coefficients have two goals: (a) they minimize deviations between predicted and actual values of the predictor variables and (b) they optimize the correlation between the predicted and obtained values of the predictor variables.

There are three types of regression techniques: standard (simultaneous) regression, sequential (hierarchical) regression, and statistical (stepwise) regression. These regression techniques mainly differ on how the predictor variables enter the equation. Before discussing the different types of multiple regression, a review of the conditions surrounding the use of multiple regression is warranted.

Conditions

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, as cited in Tabachnick & Fidell) provided a simple rule of thumb for testing multiple correlations and individual predictors in regression equations that assume a medium-size relationship between the predictor and criterion variables. Green suggested using the formula $N \geq 50 + 8m$ (where m is the number of predictors) for testing multiple correlations and $N \geq 104 + m$ for testing individual predictors. Using this rule, 130 participants would be 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.; 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. The effect size in the population can be determined in one of three ways. First, researchers can use the results of past studies to select an appropriate effect size. Second, researchers can determine that there is a minimum effect size needed for practical or theoretical significance. Third, effect sizes can be determined using conventional definitions. Cohen

(1988) has designated the R^2 values of .02, .15, and .35 for small, medium, and large effect sizes in regression analyses.

In the present study, a computer program for power analysis (Faul & Erdfelder, 1992) was used to determine if 447 participants was enough to achieve power of .80 to detect a small effect size at the .05 level of significance. Most researchers choose a power value in the range of .70 to .90 (Cohen et al., 2003). Cohen (1988) suggested that a power value of .80 is reasonable to use when there is no other basis for setting a higher or lower power. Higher power values, especially those over .90, demand large sample size that are often impractical to obtain and may detect trivial results (Judd & McClelland, 1989). The effect size of .08 was calculated by examining past research in CBM writing and determining the smallest effect size needed to differentiate between the two CBM indices that had the highest average correlations with criterion variables. Results of the power analysis revealed that a sample size of 213 would be appropriate for the present study; thus, a sample size of 447 exceeds the above specifications.

Absence of Outliers

An outlier is an extreme case that does not fit with the majority of the data. It can be a univariate outlier, an extreme score on one variable, or a multivariate outlier, an extreme combination of scores on two or more variables (Tabachnick & Fidell, 2001). Outliers may represent data that are contaminated in some way (e.g., incorrect data entry, failure to specify missing value codes, outlier not a member of the intended population) or they may be an accurate observation of a rare case (Cohen et al., 2003; Tabachnick & Fidell). Whatever the source of outliers, they can affect the estimates of regression coefficients and their standard errors, thereby contaminating the regression solution.

In addition to analyzing data for outliers by visually inspecting graphs and frequency runs, it is recommended that specialized statistics, known as regression diagnostics, be used to detect outliers (Cohen et al., 2003; Tabachnick & Fidell, 2001). Regression diagnostic statistics examine three characteristics of data points: leverage, discrepancy, and influence. Leverage tells us how far each specific data point is from the mean of the set of the predictor variables. Data points further from the mean have a greater influence on the regression model. The leverage statistic ranges from zero (no influence on the regression model) to one (highly influences the regression model). Cohen and colleagues recommended identifying a very small number of cases that have the highest leverage points and checking these cases for accuracy.

Discrepancy is the second statistic provided by the regression diagnostics. Discrepancy measures the distance between the predicted and observed values on the criterion variable (Cohen et al., 2003). Although the raw residuals provide a measure of this discrepancy, the regression line is influenced by the outlier. In other words, the outlier has pulled the regression line toward it to improve the overall fit. Therefore, other discrepancy statistics less influenced by the outliers need to be examined. Two types of discrepancy statistics are internally studentized residuals and externally studentized residuals. Externally studentized residuals are preferred over internally studentized residuals (Cohen et al.) and thus will be examined in the present study.

Externally studentized residuals examine the predicted-observed score discrepancies by deleting the outlier from the data set and recalculating the regression equation based on the remaining cases. Both large positive and negative discrepancy values indicate outliers. Cohen et al. (2003), based on the suggestions of other

researchers, recommended a cutoff value of ± 2.0. However, using this cutoff in large samples can result in an excessive number of cases to examine even if there are no real outliers in the data. Consequently, many researchers use higher cutoff scores of ± 3.0 or ± 4.0 for larger samples. Thus, in the present study the cases that have discrepancy values greater than ± 4.0 will be examined for outliers.

Influence is the last statistic provided by the regression diagnostics. Influence is the product of leverage and discrepancy (Cohen et al., 2003; Tabachnick & Fidell, 2001). It measures the change in regression coefficients if the case with the outlier was deleted from the data set. Two types of measures of influence should be examined: global measures of influence (DFFITS, Cook's D) and specific measures of influence (DFBETAS). However, if the interest of the study is overall prediction and no predictions have been made about specific regression coefficients then examining only global measures of influence are suitable for the study (Cohen et al.).

The DFFITS (i.e., difference in fit, standardized) and Cook's D provide redundant information regarding the global influence of the outlier. Both techniques compare aspects of the regression equation when the outlier case is included versus from when it is deleted from the data set. Thus, the use of one global measure is suitable for analysis (Cohen et al., 2003). The present study will examine Cook's D. The Cook's D compares the predicted values with the outlier included and deleted for all cases in the data set. The Cook's D statistic ranges from zero, no global influence, to higher numbers indicating global influence on the regression analyses. Cohen et al. recommend a cutoff of 1 or using a formula for the degrees of freedom to arrive at a critical value of the *F* distribution.

If an outlier is identified, it is recommended (Cohen et al., 2003; Tabachnick & Fidell, 2001) that the entered data be checked for accuracy. If the outlier is valid and represents a rare case, three approaches could be used. First the cases with outliers can be deleted and the data set can be reanalyzed. Second, the individual variables with the outliers can be transformed so that the data are appropriate for linear regression. Third, robust approaches that use ordinary least square methods to estimate regression coefficients can be used (Cohen et al.; Tabachnick & Fidell). Along with inspecting the frequency runs, the leverage, discrepancy, and influence of the cases will be examined in the present study. If cases with extreme outliers are detected they will be deleted from the analyses. They will not be transformed, since teachers who use CBM writing indices would be very unlikely to transform the scores. And robust approaches will not be used because many statistical programs do not include robust estimators (Cohen et al.).

Absence of Multicollinearity

In multiple regression we assume 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 (Afifi & Clark, 1997; Cohen et al.; Tabachnick & Fidell, 2001).

Screening for multicollinearity is crucial since it can cause both logical and statistical problems. 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. Although there is no accepted rule of thumb of what constitutes a high squared correlation, Tabachnick and Fidell (2001) suggested that statistical problems occur at squared correlations at or above .90.

Another way to screen for multicollinearity is by examining the variance inflation factor (VIF). The VIF is "an index of the amount that the variance of each regression coefficient is increased relative to a situation in which all of the predictor variables are uncorrelated" (Cohen et al., 2003, p. 423). If the predictor variables are uncorrelated, then the VIF is equal to 1.0. The square root of the VIF represents the impact of collinearity on the size of the standard errors for the partial regression. A common rule of thumb is that any VIF of 10 or more provides evidence of multicollinearity. Thus, a VIF of 10 means that the standard error of the partial regression coefficient is three times ($\sqrt{10}$ = 3.16) as likely to increase when compared to the situation in which all the predictors are uncorrelated. Cohen and colleagues noted that extremely high correlations between predictors are necessary to produce VIFs equal to or greater than 10. Therefore, they recommended using a more stringent guideline for VIF values.

The inverse of the VIF is called tolerance. Tolerance represents the amount of variance in one predictor variable that is independent of the other predictor variables (Cohen et al., 2003). A common rule of thumb is that tolerance values at or below .10

provide evidence of multicollinearity. A tolerance value of .10 is equal to a VIF value of 10, thus, this rule of thumb may be too lenient (Cohen et al.).

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 or creating a composite score of the variables that are highly correlated and using the composite scores in the regression analyses. For the present study, the squared correlations, the VIF values, and the tolerance values were examined to verify the absence of multicollinearity. If multicollinearity is detected through squared correlations of $\geq .90$, VIF values ≥ 10, or tolerance values $\leq .10$, then some of the predictor variables will either be omitted or a composite score among the highly correlated variables will be created.

Normality, Linearity, and Homoscedasticity of Residuals

Residuals are the differences between obtained and predicted criterion scores. There are certain assumptions made regarding the residuals when regression analysis is used. The first is that the residuals are normally distributed about the predicted criterion scores (i.e., normality). To test for normality of residuals, Cohen et al. (2003) suggested plotting a histogram of the residuals and then overlaying a normal curve with the same mean and standard deviation as the data. If the histogram and normal curve are similar, then the distribution of the residuals is normal. Additionally, a normal probability plot can be used to determine if the distribution of the residuals is normal. Second, the residuals should have a straight-line relationship with predicted criterion scores (i.e., linearity). Graphical methods are also recommended to test for linearity (Cohen et al.). The residuals can be plotted against each predictor variable and the predicted values.

Researchers can examine the graphs 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.; Tabachnick & Fidell, 2001). The same graphs used to detect linearity can also be used to detect homoscedasticity (Cohen, et al.). 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 are available to test for normality, linearity, and homoscedasticity, Cohen et al. (2003) recommended using graphical methods to help identify problems. Thus, in the present 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 will be examined to detect if there are any violations of normality, linearity, and homoscedasticity. If violations are detected, variable transformations will be considered. However, if variable transformations are deemed necessary then interpretation will be limited (Tabachnick & Fidell).

Independence of Residuals

Another assumption of regression is that the residuals of the predictions are independent of one another. Although violating this assumption does not affect the regression coefficients, it does affect the standard errors and consequently affects significance tests (Cohen et al., 2003). Scatterplots that plot residuals against sequences of cases are examined to determine if the residuals are independent. Three forms of dependency in the residuals can occur. The first form of dependency occurs when there is systematic change over time in the nature of participants or in the research procedures. Cohen et al. recommended plotting the residuals against the order of participation to

potentially reveal systematic relationships. The second form of dependency, clustering, takes place when the data are collected in groups. In this case, the errors may be more similar within the groups then between the groups. The residuals can be plotted against observation clusters to determine if residuals within groups are gathering together (Cohen et al.).

The last form of dependency, serial dependency, occurs when the data are repeatedly collected from a single individual or the same sample of individuals over time. Although the residuals can be plotted against the case numbers to detect serial dependency, a more precise statistical measure is available (Cohen et al., 2003; Tabachnick & Fidell, 2001). The statistical measure is known as autocorrelation and 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. If the coefficient is significant, it indicates nonindependence of residuals. The Durbin-Watsin coefficient will be examined in the present study to detect nonindependence of residuals. If nonindependence of residuals is discovered, data transformation will be considered (Cohen et al.).

Types of Multiple Regression

Standard Multiple Regression

The three types of multiple regression techniques differ in the procedure they use for selecting the number and order of predictor variables that enter the 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. Thus,

one predictor variable may contribute little to the prediction equation, but still be highly correlated with the criterion variable. Although there is a high correlation between this predictor and criterion variable, its unique contribution is reduced because it is correlated with another predictor variable which is accounting for the shared variability. Therefore, in standard multiple regression both the full correlation and the unique contribution of the predictor variables are considered in the interpretation of results (Tabachnick & Fidell).

Standard multiple regression assesses the relationships among predictor and criterion variables and answers two fundamental questions: (a) What is the size of the overall relationship 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 (Tabachnick & Fidell, 2001)? Therefore, the present study will use standard multiple regression to analyze the unique contributions of the predictor variables, CBM indices, to the prediction of the criterion score, the Overall Writing Quotient of the TOWL-3.

Sequential Multiple Regression

In sequential multiple regression, the researcher specifies the order in which the predictor variables are entered into the equation. The predictor variables can be entered one at a time or in blocks. Each predictor variable or set is evaluated in terms of what it adds to the equation at that point in time. The order of entry is chosen prior to data analysis and is based on logical and theoretical rationale (Tabachnick & Fidell, 2001). Cohen et al. (2003) recommended one of three basic principles that should be used in selecting the order of entry. First, the order of entry for the variables can be chosen by causal priority. The relationship between two variables may not be a direct relationship

because a third variable is the cause of both variables. Therefore, the third variable, the source of indirect relationship among the first two variables, should be entered before the two variables. As a result, "no IV entering later should be a presumptive cause of an IV that has been entered earlier" (Cohen et al., p. 158). Second, the order of entry can be determined by research relevance. In this case, the predictor variables that are of primary importance to the researcher or that have a previously established relationship with the criterion variable should be entered first (Cohen et al.). Third, researchers can examine the alternative sequences of predictor variable sets. There are times when the appropriate sequencing of predictor variables is ambiguous so alternative sequences should be analyzed and reported (Cohen et al.).

Sequential regression allows the researcher to control the number and order of entry of predictor variables in the regression equation. In sequential regression, the statistical power of a test regarding an explicit hypothesis is likely to be maximized since the results of the test are not confounded with predictor variables that are less important (Cohen et al., 2003). When sequential regression is used the fundamental question is: Does a certain predictor variable significantly add to the prediction of a criterion variable after variance due to other predictor variables are accounted for (Tabachnick & Fidell)? Because the present study does not ask this question, sequential regression will not be used except as an aid to interpret results and explore post-hoc analyses.

Statistical Regression

In statistical regression, the order of entry of predictor variables is based solely on statistical criteria. The first predictor variable that is entered into the equation is the one with the highest correlation with the dependent variable. The second predictor variable

entered is the one that results in the largest unique contribution to R^2 (the proportion of variance of the criterion variable that is explained by the predictor variables). This process is continued until there are no predictor variables left that make unique contributions at a statistically significant level specified by the researcher (Cohen et al., 2003; Tabachnick & Fidell, 2001). There are three variations of statistical regression: forward selection, backward deletion, and stepwise regression. In forward selection, predictor variables are entered one at a time and only if they meet certain statistical criteria. In backward deletion, all the predictor variables are entered at once and then they are deleted one at a time if they do not contribute significantly to the prediction of the criterion variable. And in stepwise regression, the predictor variables are added into the equation one at a time if they meet statistical criteria. The difference between stepwise and forward selection is that in stepwise regression previously entered variables can be eliminated at any step if they no longer contribute significantly to the prediction of the criterion variable (Tabachnick & Fidell).

In stepwise regression, decisions on which variables to include in the regression equation are based on minor differences in statistics computed from a single sample (Cohen, et al., 2003; Tabachnick & Fidell, 2001). Therefore, stepwise regression is not recommended because it tends to capitalize on chance and overfit data. Additionally, a larger and more representative sample is needed for statistical regression than the other types due to the above limitations (Cohen, et al.; Tabachnick & Fidell). Cohen et al. recommended that when stepwise regression is used, the following three conditions be satisfied: (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. When statistical

regression is used, the primary question being answered is: What is the best linear combination of predictor variables to predict the criterion variable for this specific sample (Tabachnick & Fidell)? Because the present study does not ask this question stepwise regression will not be used for data analysis.

Statistical Inferences

The multiple correlation coefficient (R^2) is a measure of association between the predictor variables and the criterion variable (Cohen et al., 2003). It is the proportion of variance shared between the criterion variable and the predictor variables. The overall inferential test (the *F* test) in multiple regression measures 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. However, with a large sample size, this statistic becomes trivial because the null hypothesis will most likely be rejected (Tabachnick & Fidell, 2001).

In standard multiple regression, the *t*-test is used to examine the significance of individual standardized partial regression coefficients (β). It is important to note that 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 the present study, the significance of β will be examined. However, since β only examines the unique contribution of each predictor, the correlations between each individual predictor variable and criterion variable will also be examined.

RESULTS

Preliminary Analyses

Reliability

Interrater Reliability

Student protocols were divided among the ten scorers, with each rater scoring approximately 40 CBM protocols and 40 TOWL-3 protocols. Each protocol was initially scored by the primary researcher, independently scored by a trained doctoral student in school psychology, and rescored for clerical and computational errors by the primary researcher. Thus, each protocol was scored by the primary researcher and an independent rater. Average interscorer reliability between the primary researcher and independent raters was extremely high. All correlations were above .94 for the CBM indices and .99 for the Overall Writing Quotient of the TOWL-3 (see Table 1).

Table 1

Interrater Reliability for the CBM Indices and the Overall Writing Quotient

Measures	Coefficients
Total Words Written	0.99
Words Spelled Correctly	0.99
Percentage of Words Spelled Correctly	0.96
Number of CWS	0.99
Percentage of CWS	0.95
Number of CWS-IWS	0.98
Number of Sentences	0.96
Number of Correct Capitalizations	0.99
Number of Punctuation Marks	0.98
Percentage of Correct Punctuation Marks	0.98
Overall Writing Quotient	0.99

Note. CWS = correct words sequences; CWS-IWS = correct word sequences minus incorrect word sequences.

Internal Consistency Reliability

Cronbach's coefficient alpha was used to determine the degree of homogeneity among the test items of the TOWL-3. Alpha coefficients were calculated for the scores obtained from the Overall Writing Quotient, the Contrived Writing Quotient, and the Spontaneous Writing Quotient from both independent raters. Although the coefficient alphas for the scores obtained from the Overall Writing Quotient fell below that of the

test authors' research (Hammill & Larsen, 1996), it was above the acceptable range of .80 (Salvia & Ysseldyke, 2001). However, the alpha coefficients for the Contrived and Spontaneous Writing Quotients were below those found in the normative sample and what is recommended for screening and decision making (Salvia & Ysseldyke; see Table 2). Thus, analyses were conducted with only the more reliable Overall Writing Quotient.

Table 2

Internal Consistency Reliability Coefficients for the TOWL-3 Scores

	Internal Consistency Reliability Coefficients	
Quotient	Rater One	Rater Two
Overall Writing Quotient	.84	.85
Contrived Writing Quotient	.78	.79
Spontaneous Writing Quotient	.73	.73

Conditions

Prior to analyses, all ten CBM indices and the Overall Writing Quotient were examined through various scatterplots and statistical equations using SPSS programs (version 10.0 for Windows) for accuracy of data entry, absence of outliers, absence of multicollinearity, and fit between their distributions and the assumptions of multivariate analysis. Results of the regression diagnostics, which included leverage, discrepancy, and influence, revealed five outliers. Four of the outliers were caused by a clerical error during data entry and were reentered correctly. The other outlier was from a student in special education whose scores were entered correctly. Thus, the individual's scores remained unchanged in the data sample. The regression diagnostics were run again after the clerical errors were corrected. 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 statistic were all below one. Therefore, no extreme outliers were detected or deleted from the analysis.

In addition to exploring outliers, the data was screened for multicollinearity. When examining all 10 CBM indices as independent variables, multicollinearity was present as evident through squared correlations of $\geq .90$, VIF values ≥ 10, or tolerance values $\leq .10$. Therefore, to reduce estimation problems as a result of the redundancy among predictor variables, three of the predictor variables (i.e., words spelled correctly, correct word sequences, and total punctuation marks) were omitted from the study (Cohen et al., 2003; Morrow-Howell, 1994). These three predictor variables were chosen based on their high correlations with other predictor variables and past research showing more empirical support for the use of accuracy-based measures (i.e., percentage of WSC and percentage of CWS) than fluency-based measures (i.e., WSC and CWS) with secondary students (Jewell & Malecki, 2005; Tindal & Parker, 1989, 1991; Watkinson & Lee, 1992). Before omitting the predictor variables, linear transformations of the variables were attempted; however, multicollinearity continued to exist. Once the three predictor variables were removed from the analyses, all squared correlations were below .90. However, two of the seven remaining predictor variables (i.e., total words written and correct minus incorrect word sequences) had tolerance values below .10. and VIF values greater than 10. These two predictor variables were retained in the analyses due to past research determining the relevance of these two indices to the theory of CBM. If these two predictor variables were removed from the study, the estimates of all the other predictor variables may be biased in their absence.

Finally, 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 normal and the small skewness that appeared in the histogram should not affect conclusions. Additionally, a normal probability plot revealed a nearly straight line indicating the assumption of normally distributed residual error was met. A set of scatterplots examining the linear relationship between the predictor and criterion variables and the residuals and the predictor variables were inspected. There was no substantial departure from the assumption of linearity as seen by the lowess lines that were roughly horizontal with minor curves. However, when examining the scatterpolots of the residuals and the predictor variables for homscedasticity, the graphs revealed that both percent measures (i.e., PWSC and PCWS) did not evenly spread out. More than likely this is because percent measures are ordinal variables that often have distributions that are rectangular instead of normal (Cohen et al., 2003; Tabachnick & Fidell, 2001). Results of these graphs led to the transformation of the two percent variables to reduce skewness and kurtosis and improve homoscedasticity of residuals. When the arcsine transformation was used to normalize PWSC and PCWS, the overall results were the same except that CWS-IWS also became a significant predictor. 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. The Durbin-Watsin coefficient indicated that the residuals were independent.

Descriptive Statistics

Descriptive statistics for all curriculum-based measures and the Overall Writing Quotient are presented in Table 3.

Table 3

Descriptive Statistics on CBM Indices and the Overall Writing Quotient

Measure	*M*	*SD*	Skew	Kurtosis
TWW	59.62	16.94	0.05	0.03
WSC	58.36	17.01	0.01	0.02
%WSC	97.63	3.92	-4.56	34.25
CWS	60.29	19.36	0.06	0.01
%CWS	90.58	9.62	-2.34	8.53
CWS-IWS	54.63	20.92	-0.17	0.18
SEN	5.40	2.02	0.47	0.50
CC	6.56	3.02	0.72	1.06
TPM	8.23	3.54	0.59	0.99
CPM	7.80	3.33	0.59	1.29
Overall Writing Quotient	96.35	13.51	-0.64	0.67

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. CWS-IWS = number of correct word sequences minus incorrect word sequences. SEN = number of sentences. CC = number of correct capitalizations. TPM = number of punctuation marks. CPM = number of correct punctuation marks.

Relationship Between the Overall Writing Quotient and CBM Indices

Standard regression analysis were conducted to explore which CBM indices best predicted the TOWL-3 Overall Writing Quotient. Students' scores on total words written, percentage of words spelled correctly, percentage of correct word sequences, correct minus incorrect word sequences, number of sentences, number of correct capitalization, and number of correct punctuation marks were entered simultaneously as predictors in a regression analysis that included students' scores on the TOWL-3 Overall Writing Quotient as the dependent variable. Intercorrelations among the CBM indices and the Overall Writing Quotient are presented in Table 4 and results of the regression analysis are shown in Table 5.

Table 4

Intercorrelations Among the CBM Indices and the Overall Writing Quotient

	TWW	WSC	%WSC	CWS	%CWS	CWS-IWS	SEN	CC	TPM	CPM	Overall Writing Quotient
TWW	1.000	.994	.222	.949	.272	.857	.655	.525	.578	.587	.338
WSC		1.000	.315	.964	.337	.888	.654	.523	.588	.598	.371
%WSC			1.000	.386	.740	.491	.191	.144	.255	.268	.414
CWS				1.000	.510	.971	.683	.574	.686	.701	.485
%CWS					1.000	.677	.232	.222	.326	.362	.606
CWS-IWS						1.000	.624	.530	.653	.677	.561
SEN							1.000	.677	.711	.726	.277
CC								1.000	.663	.670	.225
TPM									1.000	.979	.429
CPM										1.000	.443
Overall Writing Quotient											1.00

Note. All correlations are significant ≥ .01.

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. CWS-IWS = number of correct word sequences minus incorrect word sequences. SEN = number of sentences. CC = number of correct capitalizations. TPM = number of punctuation marks. CPM = number of correct punctuation marks.

Table 5

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

Variable	*B*	*SE B*	β	*t*	*p*
Intercept	49.301	15.766			
TWW	0.018	.101	.022	0.176	.860
%WSC	-0.257	.186	-.074	-1.379	.169
%CWS	0.689	.143	.490	4.813	.001
CWS-IWS	0.094	.113	.146	0.836	.403
SEN	-0.446	.404	-.067	-1.103	.270
CC	-0.521	.233	-.117	-2.237	.026
CPM	1.199	.254	.296	4.725	.001

Note. TWW = total words written. %WSC = percentage of words spelled correctly. %CWS = percent of correct word sequences. CWS-IWS = correct word sequences minus incorrect word sequences. SEN = number of sentences. CC = number of correct capitalizations. CPM = number of correct punctuation marks.

The multiple correlation coefficient was significantly different from zero and the seven predictor variables collectively accounted for 44% of the variance in students' Overall Writing Quotient scores, F (7, 439) = 49.40, $p < .001$. Only three of the predictor variables significantly contributed unique variance to the prediction of the Overall Writing Quotient scores. As seen in Table 5, percentage of CWS contributed the most unique variance (β = .490, $p < .01$), followed by the number of correct punctuation marks (β = .296, $p < .01$) and the number of correct capitalizations (β = -.117, $p < .05$).

Post Hoc Analyses

Due to the high correlation between total words written and correct minus incorrect word sequences (r = .86), these two variables were individually removed for post hoc analyses. After removing total words written, the first analysis included students' scores on percentage of words spelled correctly, percentage of correct word sequences, correct minus incorrect word sequences, number of sentences, number of correct capitalization, and number of correct punctuation marks as simultaneous predictors in a regression analysis that included students' scores on the TOWL-3 Overall Writing Quotient as the dependent variable. The multiple correlation coefficient was significantly different from zero and the six predictor variables collectively accounted for 44% of the variance in students' Overall Writing Quotient scores, F (6, 440) = 57.75, p < .001.

Without total words written in the regression analysis, four of the predictor variables significantly contributed unique variance to the prediction of the Overall Writing Quotient scores. As seen in Table 6, percentage of correct word sequences and correct punctuation mark contributed the most unique variance, respectively (β = .447, p < .01; β = .293, p < .01), followed by correct word sequences minus incorrect word sequences (β = .174, p < .05) and the number of correct capitalizations (β = -.116, p < .05).

Table 6

Summary of CBM Index Scores, excluding Total Words Written, as Predictors of Overall Writing Quotient Scores

Variable	*B*	*SE B*	β	*T*	*p*
Intercept	50.587	13.965			
%WSC	-0.251	.183	-.073	-1.371	.171
%CWS	0.669	.092	.477	7.264	.001
CWS-IWS	0.113	.044	.174	2.583	.010
SEN	-0.434	.398	-.065	-1.090	.276
CC	-0.520	.233	-.116	-2.235	.026
CPM	1.186	.243	.293	4.884	.001

Note. %WSC = percentage of words spelled correctly. %CWS = percent of correct words sequences. CWS-IWS = correct word sequences minus incorrect word sequences. SEN = number of sentences. CC = number of correct capitalizations. CPM = number of correct punctuation marks.

Next, students' scores on correct word sequences minus incorrect word sequences were removed from the regression analysis. Thus, students' scores on total words written, percentage of words spelled correctly, percentage of correct word sequences, number of sentences, number of correct capitalization, and number of correct punctuation marks were entered simultaneously as predictors in the regression analysis. The multiple correlation coefficient was significantly different from zero and the six predictor variables collectively accounted for 44% of the variance in students' Overall Writing Quotient scores, F (6, 440) = 57.55, $p < .001$.

Without correct word sequences minus incorrect word sequences in the regression analysis, four of the predictor variables significantly contributed unique variance to the prediction of the Overall Writing Quotient scores. As seen in Table 7, percentage of

correct word sequences and correct punctuation mark contributed the most unique variance, respectively (β = .562, p < .01; β = .315, p < .01), followed by total words written (β = .119, p < .05) and the number of correct capitalizations (β = -.117, p < .05). Based on the two regression analyses, it appears that total words written and correct minus incorrect word sequences are redundant variables in the prediction of Overall Writing Quotient scores. However, past research has shown that correct minus incorrect word sequences better predicts criterion measures in written expression for secondary students than does total words written (Espin et al., 2000; Weissenburger & Espin, 2005). Thus, correct minus incorrect word sequences may be a more valid CBM index for measuring eighth-grade students' writing ability.

Table 7

Summary of CBM Index Scores, excluding Correct Words Minus Incorrect Words, as Predictors of Overall Writing Quotient Scores

Variable	*B*	*SE B*	β	*T*	*p*
Intercept	42.772	13.692			
TWW	-0.095	.039	.119	2.448	.015
%WSC	-0.283	.183	.082	-1.546	.123
%CWS	0.789	.077	.562	10.218	.001
SEN	-0.455	.404	-.068	-1.128	.260
CC	-0.523	.233	-.117	-2.246	.025
CPM	1.277	.236	.315	5.409	.001

Note. TWW = total words written. %WSC = percentage of words spelled correctly. %CWS = percent of correct words sequences. SEN = number of sentences. CC = number of correct capitalizations. CPM = number of correct punctuation marks.

In addition to the above standard multiple regression analyses, sequential multiple regression analyses were used to help interpret results and explore post hoc analyses. Sequential regression was used to determine if the addition of correct punctuation marks and all other CBM indices improved prediction of the Overall Writing Quotient scores beyond that afforded by percent of correct word sequences. Table 8 provides a summary of the results. The multiple correlation coefficient was significantly different from zero at the end of each step. After step 1, with percent of correct word sequences in the equation, $R^2 = .37$, $F\ (1, 445) = 258.19$, $p < .001$. After step 2, with correct punctuation marks added to the prediction of Overall Writing Quotient scores, $R^2 = .42$, $F_{inc}\ (1, 444) = 42.69$, $p < .001$. After step 3, with all other CBM indices added to the prediction of Overall Writing Quotient scores, $R^2 = .44$, $F_{inc}\ (8, 436) = 2.09$, $p < .05$. Moreover, the

semipartial correlation at each step ($\sqrt{\Delta R^2}$ = .61, .24, and .15, respectively) was statistically and educationally significant. According to Hunsley and Meyer (2003), a semiparital correlation of above .15 at step 3 would indicate a reasonable contribution to the regression equation. However, at the third step only correct capitalizations was statistically significant (β = -.123, $p < .05$) and was more than likely being suppressed by another variable as evident by its negative sign of the regression weight and positive correlation (r = .23; Cohen et al., 2003; Tabachnick & Fidell, 2001). Thus, correct punctuation marks was the only other CBM index to have incremental predictive validity in the prediction of Overall Writing Quotient scores beyond that afforded by percent of correct word sequences.

Table 8

Summary of Sequential Regression Analysis for Variables Predicting Overall Writing Quotient Scores with the Percent of Correct Word Sequences Entered First

Variable	*B*	*SE B*	β	*T*	*p*
Step 1					
%CWS	0.851	.053	.606	16.068	.001
Step 2					
%CWS	0.723	.054	.515	13.305	.001
CPM	1.024	.157	.253	6.534	.001
Step 3					
%CWS	0.860	.223	.613	3.858	.001
CPM	0.291	.747	.072	.389	.697
TWW	-0.890	1.082	-1.116	-.822	.411
WSC	0.834	1.092	1.051	.764	.445
%WSC	-0.700	.555	-.203	-1.261	.208
CWS	0.401	.576	.575	.697	.486
CWS-IWS	-0.242	.410	-.375	-.589	.556
SEN	-0.410	.405	-.061	-1.012	.312
CC	-0.551	.235	-.123	-2.339	.020
TPM	0.798	.697	.209	1.144	.253

Note. %CWS = percent of correct words sequences. CPM = number of correct punctuation marks. TWW = total words written. WSC = number of words spelled correctly. %WSC = percentage of words spelled correctly. CWS = number of correct word sequences. CWS-IWS = number of correct word sequences minus incorrect word sequences. SEN = number of sentences. CC = number of correct capitalizations. TPM = number of punctuation marks.

A second sequential regression was employed to determine if the addition of percent of correct word sequences and all other CBM indices improved prediction of Overall Writing Quotient scores beyond that afforded by correct punctuation marks. Table 9 provides a summary of the results. The multiple correlation coefficient was significantly different from zero at the end of each step. After step 1, with correct punctuation marks in the equation, only 19% of the variability was accounted for in the prediction of Overall Writing Quotient scores, F (1, 445) = 106.09, $p < .001$. After step 2, with percent of correct word sequences added to the prediction of Overall Writing Quotient scores, 42% of the variability was accounted for in the prediction of Overall Writing Quotient scores, F_{inc} (1, 444) = 177.02, $p < .01$. After step 3, with all other CBM indices added to the prediction of Overall Writing Quotient scores, $R^2 = .44$, F_{inc} (8, 436) = 2.09, $p < .05$. Similar to the preceding regression equation, the semipartial correlation at each step was statistically and educationally significant as outlined by Hunsley and Meyer (2003; $\sqrt{\Delta R^2} = .44$, .48, and .15, respectively). However, at step 3, correct capitalization was the only predictor variable that was statistically significant ($\beta = -.123$, $p < .05$). As previously mentioned, another variable was apparently suppressing the value of correct capitalizations. Therefore, the percentage of correct word sequences was the only predictor to add incremental validity in the prediction of the Overall Writing Quotient scores beyond that afforded by correct punctuation marks.

Table 9

Summary of Sequential Regression Analysis for Variables Predicting Overall Writing Quotient Scores with Correct Punctuation Marks Entered First

Variable	*B*	*SE B*	β	*T*	*p*
Step 1					
CPM	1.777	.173	.439	10.300	.001
Step 2					
CPM	1.024	.157	.253	6.534	.001
%CWS	0.723	.054	.515	13.305	.001
Step 3					
CPM	0.291	.747	.072	.389	.697
%CWS	.860	.223	.613	3.858	.001
TWW	-0.890	1.082	-1.116	-.822	.411
WSC	0.834	1.092	1.051	.764	.445
%WSC	-0.700	.555	-.203	-1.261	.208
CWS	0.401	.576	.575	.697	.486
CWS-IWS	-0.242	.410	-.375	-.589	.556
SEN	-0.410	.405	-.061	-1.012	.312
CC	-0.551	.235	-.123	-2.339	.020
TPM	0.798	.697	.209	1.144	.253

Note. CPM = number of correct punctuation marks. %CWS = percent of correct words sequences. TWW = total words written. WSC = number of words spelled correctly. %WSC = percentage of words spelled correctly. CWS = number of correct word sequences. CWS-IWS = number of correct word sequences minus incorrect word sequences. SEN = number of sentences. CC = number of correct capitalizations. TPM = number of punctuation marks.

The above analyses demonstrated that percent of correct word sequences and correct punctuation marks are the most parsimonious model in predicting Overall Writing Quotient scores for eighth-grade students. Moreover, as seen in step 1 of the sequential analyses, percentage of correct word sequences ($R^2 = .37$) accounted for more variance in predicting Overall Writing Quotient scores than correct punctuation marks ($R^2 = .19$). Thus, percentage of correct word sequences appears to be the best predictor of Overall Writing Quotient scores in a sample of eighth-grade students.

DISCUSSION

Assessment measures are commonly used to make important decisions regarding students' lives. Although there are several measures available for assessing writing skills, many are subjective, difficult to score, and time-consuming (Watkinson & Lee, 1992). For these reasons, researchers continue to explore more straightforward, time-efficient, and informative methods of writing assessment, such as CBM writing indices. If findings support the use of curriculum-based scoring indices for writing, school personnel can utilize these scores for assessing and monitoring progress of students' writing skills. Thus, criterion-related evidence for the validity of CBM writing indices was gathered by examining the ability of these indices to predict scores from a well designed measure of written expression, the TOWL-3.

Past research has suggested that CBM indices need to measure more complex skills in older students' writing than the traditional total words written. However, as indices become more complex, interscorer reliability may tend to decrease. In the present study, all CBM indices were scored reliably following training. Average interscorer reliability between the primary researcher and independent raters was extremely high and above .94 for the CBM indices, regardless of the level of complexity involved in scoring different CBM indices. Additionally, the interscorer reliability between the primary researcher and independent raters was .99 for the Overall Writing Quotient. However, it's important to note that although interscorer reliability was high, the principal researcher provided extensive training as well as monitoring throughout the scoring period to obtain consistency among raters. As the indices become more complex, more time is needed in training and monitoring of scorers and in scoring the index itself. Additionally, there is

extensive research that has consistently found a high incidence of clerical errors on psychological tests, including cognitive and objective personality tests (Allard & Faust, 2000; Charter, Walden, & Padilla, 2000; Sullivan, 2000). These simple errors can result in significant changes in standardized scores. Thus, the principal examiner rechecked all protocols for clerical and computational errors (e.g., incorrect addition of raw scores, incorrect discontinuation rules, etc.) to ensure the validity of test scores.

The results of the present study provide only modest support for the use of CBM indices in written expression. The seven CBM indices accounted for only 44% of the variance in TOWL-3 scores. Moreover, only three of the seven predictors uniquely contributed to the prediction of TOWL-3 scores. Two of the three variables, percent of correct word sequences and correct punctuation marks, appear to be potential candidates for use as CBM indices at the secondary level.

Past research has shown the percentage of CWS to serve as a moderate to good indicator of writing performance for secondary students (Jewell & Malecki, 2005; Parker, Tindal, & Hasbrouck, 1991; Tindal & Parker, 1989; Watkinson & Lee, 1992). Tindal and Parker's initial factor analytic study had suggested that the percentage of CWS represents a production-independent factor that measures accuracy and not fluency. As students' age increases, measures of writing accuracy may be more strongly related to students' performance on writing criteria than measures of writing fluency (Jewell & Malecki; Parker, Tindal, & Hasbrouck; Tindal & Parker; Watkinson & Lee). This is consistent with the present study in which the accuracy based index, percentage of CWS, had the strongest bivariate correlation with TOWL-3 scores ($r = .61$) and contributed the most unique variance in the prediction of TOWL-3 scores ($\beta = .49$).

Although correct punctuation marks added incremental validity in the prediction of TOWL-3 scores, it accounted for less variance than percentage of CWS. As seen in step 1 of the sequential regression models, correct punctuation marks only accounted for 19% of the variance while percentage of CWS accounted for 37% of the variance in TOWL-3 scores.

Although this study, along with past research, found the percentage of CWS to be a fair indicator of writing performance (Jewell & Malecki, 2005; Parker, Tindal, & Hasbrouck, 1991; Tindal & Parker, 1989; Watkinson & Lee, 1992) this index should be used with caution for progress monitoring (Espin et al., 2000; Parker, Tindal, & Hasbrouck 1991; Tindal & Parker, 1989). Percentage measures, including the percentage of CWS, are linear transformations of raw scores. Movement across percentage values depends on the number of opportunities the student has to respond. Thus, students' progress may be masked depending on the number of CWS the student has written. For example, a student can write 40 word sequences with 38 correct in the fall and 80 word sequences with 70 correct in the spring. Although this student's number of CWS has increased from 40 to 80, the percentage of CWS has decreased from 95% to 88%.

In addition to the percentage of CWS, a related index, CWS-IWS, has been shown to be a reliable and valid indicator of writing performance (Espin et al., 2000; Espin et al., 2005; Malecki & Jewell, 2003; Weissenburger & Espin, 2005). In the current study, CWS-IWS did not significantly contribute unique variance in the prediction of TOWL-3 scores in the initial data analysis. However, CWS-IWS and TWW were highly correlated (r = .86); therefore contributing less unique information to the prediction of TOWL-3 scores. Thus, a post hoc analysis was completed with the removal of TWW from the

regression equation. Once TWW was removed from the equation CWS-IWS was a significant predictor of TOWL-3 scores (β = .17). Additionally, CWS-IWS was moderately correlated with TOWL-3 scores (r = .56). These results are consistent with past research which has shown CWS-IWS to have moderate to strong correlations with different writing criteria, including teachers' ratings of writing proficiency, statewide testing, and holistic ratings of writing (Espin et al., 2000; Weissenburger & Espin, 2005). In addition to criterion-related validity, CWS-IWS has shown to be a reliable index (Espin et al., 2000; Espin et al., 2005) and one that is sensitive to change over time (Espin et al., 2005; Malecki & Jewell, 2003). Although there is evidence that CWS-IWS is more technically sound for older students, it did not significantly predict the Overall Writing Quotient scores for this eighth-grade sample in the initial data analysis. Furthermore, in the sequential post hoc analyses it did not add incremental validity beyond that afforded by percentage of CWS and correct punctuation marks in predicting TOWL-3 scores. Thus, without further research adding to the validity of CWS-IWS, this index should not be used to predict writing ability for eighth grade students.

The second CBM scoring index that significantly contributed unique variance to the prediction of TOWL-3 scores was correct punctuation marks (β = .30). The correlation between correct punctuation marks and TOWL-3 scores was moderate (r = .44) and similar to past studies examining the relationship between correct punctuation marks and standardized writing assessments (Gansle et al., 2002; Gansle et al., 2006). Based on the current study, the number of correct punctuation marks may be a usable index of writing skill. However, a very limited number of studies have examined the use of correct punctuation marks (Gansle et al., 2002; Gansle et al., 2004; Gansle et al.,

2006). Additionally, the primary researchers of these studies are the same and results of these studies have been inconsistent. Although the present study and Gansle and colleagues' (2002) research found that the number of correct punctuation marks significantly contributed to the prediction of a writing criterion, Gansle and colleagues' (2004) study found that the number of correct punctuation marks did not significantly contribute to the prediction of WJ-R writing samples. However, given the good interscorer agreement and its relative ease in scoring, correct punctuation marks may be a promising CBM index if more research suggests it's a reliable and valid indicator of writing.

The last CBM index to significantly contribute to the prediction of TOWL-3 scores was correct capitalizations. However, this needs to be viewed with caution. It is probable that another predictor variable was a negative suppressor for correct capitalizations. Thus, the relationship between one of the predictor variables and correct capitalization is hiding the real relationship between correct capitalization and TOWL-3 scores. This is largely evident by the negative sign of the regression weight of correct capitalization (β = -.12) which is opposite of what would be expected since correct capitalization and TOWL-3 scores are positively correlated (r = .23; Cohen et al., 2003; Tabachnick & Fidell, 2001). Unfortunately, it's difficult to determine which predictor variable is doing the suppression since the study includes several CBM indices in the analysis. It's important to note that the weak bivariate correlation between correct capitalization and TOWL-3 scores was consistent with past research, indicating that correct capitalization may not be a valid indicator of written expression as measured by different writing criteria. Also, the number of correct capitalization can be highly

dependent on what students write. For example, one student may choose to write a story referring to several individuals or places using proper names; while another student may write about an event that does not provide opportunities for correct capitalizations (Gansle et al., 2006). Therefore, the number of correct capitalizations should not be used to measure students' writing ability, unless future research finds additional evidence supporting its use.

Comparing the CBM literature

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. More notably, it has been concluded that the research data supports using oral reading fluency as an indicator of performance and progress in reading for elementary school students. The oral reading fluency 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 et al., 2007). Unfortunately, the same can not be said for CBM in math and writing. In these two areas, there is not one CBM index that has been consistently shown to be technically sound or theoretically appropriate (Foegen et al., 2007; McCaster & Espin, 2007). Moreover, in CBM writing recent research has yielded only moderate criterion validity at best and these coefficients have been more modest than those obtained in other areas of CBM research (McCaster & Espin, 2007). For example, the present study yielded weak (r = .26) to moderate (r = .61) validity coefficients between CBM writing indices and the TOWL-3. In contrast, reading CBM validity coefficients have consistently been in the .70s or above (Wayman et al.).

Additionally, in all three subject areas most research has been conducted at the elementary level. Specific to CBM in writing, past research has reported that the criterion-related validity evidence decreases as a student gets older (Jewell & Malecki, 2005; Weissenburger & Espin, 2005). For example, Weissenburger and Espin found that criterion validity coefficients between a statewide test and CBM writing indices (i.e., TWW, CWS, and CWS-IWS) decreased with age. For fourth-grade students the criterion validity coefficients ranged from .36 to .68. For students in the eighth grade, the validity coefficients were lower, ranging from .24 to .63. More remarkably, validity coefficients for tenth-grade students were weak and ranged from .04 to .36. Students' writing tend to become more complex and multifaceted with age. Thus, secondary students' writing appears to be too complex for a single metric to validly measure. Research has consistently shown that accuracy measures, along with more complex scoring measures, have stronger technical characteristics than do simple fluency measures when measuring secondary students' writing ability. This was consistent with the present study in which the percentage of CWS, an accuracy-based metric, had the strongest correlation with TOWL-3 scores. Moreover, it accounted for the most variance in predicting TOWL-3 scores among this sample of eighth-grade students.

Limitations and Future Research

Although the above results provide evidence for the possible use of two CBM writing indices, the percentage of CWS and correct punctuation marks, there are limitations of this study that require discussion. First, multicollinearity was present among the 10 CBM indices that were collected from the writing samples. To reduce estimation problems, three of the predictor variables (i.e., words spelled correctly, correct

word sequences, and total punctuation marks) were omitted from the study. These three predictor variables were chosen based on their high correlations with other predictor variables and past research. Secondary school studies have shown more empirical support for the use of accuracy-based measures (i.e., percentage of WSC and percentage of CWS) than fluency-based measures (i.e., WSC and CWS) (Jewell & Malecki, 2005; Tindal & Parker, 1989, 1991; Watkinson & Lee, 1992). Thus, the two fluency-based measures, WSC and CWS, were removed from the analyses and their related indices, percentage of WSC and percentage of CWS, were retained in the analyses. As for total punctuation marks, this was removed due to the practicality that its related index, correct punctuation marks, takes into consideration the quality of student writing.

After the removal of these three predictor variables, TWW and CWS-IWS had the highest squared correlation ($r^2 = .74$). Additionally, TWW and CWS-IWS had tolerance values below .10 and VIF values greater than ten. Although there was a possibility that high multicollinearity will affect the interpretation of the regression coefficients, these two variables were retained in the initial analysis due to their relevance as outlined in past research. Although TWW has not been viewed as a valid indicator of older students' written expression skills (Espin et al., 2000; Espin, et al., 1999; Gansle et al., 2002; Malecki & Jewell, 2003; Parker, Tindal, & Hasbrouck, 1991; Watkinson & Lee, 1992; Weissenburger & Espin, 2005), it has consistently been shown to be a valid and reliable estimate of writing ability for elementary students (Deno, Marston, & Mirkin, 1982; Espin et al., 1999; Espin et al., 2000; Gansle et al, 2004; Marston et al, 1981). Additionally, it is the CBM writing index that most educators are familiar with. CWS-IWS was retained in the initial data analysis due to the growing popularity of this index

as a method of assessing secondary students' writing ability (Espin et al., 2000; Espin et al., 2005; Malecki & Jewell; Weissenburger & Espin). However, due to the likelihood of high multicollinearity between these two predictor variables, they were individually removed from post hoc analyses. After removing TWW from the data analysis, CWS-IWS significantly contributed to the prediction of TOWL-3 scores (β = .17). Also, TWW was a significant predictor of TOWL-3 scores (β = .12) after CWS-IWS was removed from the regression analysis. Based on the two regression analyses, it appears that TWW and CWS-IWS are redundant variables in the prediction of TOWL-3 scores. It is important to note that neither variable added incremental validity in the prediction of TOWL-3 scores beyond that afforded by the percentage of CWS and correct punctuation marks.

Another method that can be used to address multicollinearity is to create a composite score of the variables that are highly correlated (Cohen et al., 2003; Morrow-Howell, 1994; Tabachnick & Fidell, 2001). One way to create a composite score is through a principal components analysis. This was considered and rejected because educators, specifically teachers, would be unlikely to conduct a principal components analysis of their data prior to analyzing the scores. Research must be responsive to the needs of teachers and educators. Often, teachers' interests and actions are constrained by other aspects of instruction (i.e., working with large groups, time, experience level, education level, etc.). Thus, a portion of research must examine practices within a context that is useful for teachers (Cannon, 2006; Lloyd, Weintraub, & Safer, 1997). Additionally, adding a principal components analysis is not adhering to CBM's model of

formative assessment in which educators can produce and score CBM probes frequently and with ease (Deno, 2003; Shinn & Bamonto; Shinn, Rosenfield, & Knutson, 1989).

A second limitation to the study is that all students who participated were from one grade level in a single school district in Central New Jersey. Thus, generalizations beyond this sample should be limited. Replication of this study with students from different grade levels and from different regions of the country would allow for a more constructive examination of the research question raised in this study. Specifically, future research needs to explore grade-level trends to further delineate which CBM indices are most appropriate to use with different grade levels.

Finally, this study did not investigate the suitability of any of the 10 CBM indices for progress monitoring, program evaluation, eligibility decisions, or skill diagnosis. In the current study, both the CBM and criterion measure were given at one point in time, only providing evidence of criterion-related validity. Future research needs to examine the utility of CBM indices for a variety of academic purposes. Additionally, when research examines the application of CBM indices, different grade levels should be used to further determine the reliability and validity of CBM indices in written expression.

Conclusion

The present study examined the relationship between CBM scores and a well-designed standardized test of writing, the TOWL-3. Results of this study were consistent with past research that shows simple fluency measures, such as TWW, are not sufficient for assessing secondary students' writing ability. However, results indicated that a more complex fluency measure, the number of correct punctuation marks, and an accuracy-based measure, the percentage of CWS, were the best predictors of TOWL-3 scores for

eighth-grade students. Of the two CBM indices, the percentage of CWS contributed the most unique variance in predicting TOWL-3 scores and had the strongest bivariate correlation with TOWL-3 scores. These results are consistent with previous research on the use of percentage of CWS as a general indicator of writing ability for secondary students (Jewell & Malecki, 2005; Parker, Tindal, & Hasbrouck, 1991; Tindal & Parker, 1989; Watkinson & Lee, 1992). However, it is important to note that the percentage of CWS only predicted 37% of the variance in TOWL-3 scores. Additionally, past research has not shown the percentage of CWS to be adequate for progress monitoring (Espin et al., 2000; Parker, Tindal, & Hasbrouck 1991; Tindal & Parker, 1989).

Results of the current study, along with past research, do not lend strong support to the use of CBM to assess and monitor writing skill at the secondary level. Although future research may suggest a CBM measure that yields reliable and valid scores and that can be used to monitor progress, current research suggests that existing CBM measures are unlikely to fulfill this need. Therefore, for instructional or high-stakes decisions, at this point in time educators may wish to rely on other qualitative and quantitative aspects of student writing, including published norm-referenced tests.

REFERENCES

Afifi, A. A., & Clark, V. (1997). *Computer-aided multivariate analysis* (3rd ed.). Boca Raton: Chapman & Hall/CRC.

Albin, M. L., Benton, S. L., & Khramtsova, I. (1996). Individual differences in interest and narrative writing. *Contemporary Educational Psychology, 21,* 305-324.

Algozzine, B., O'Shea, D. J., Stoddard, K., & Crews, W. G. (1988). Reading and writing competencies of adolescents with learning disabilities. *Journal of Learning Disabilities, 21,* 154-160.

Allard, G. & Faust, D. (2000). Errors in scoring objective personality tests. *Assessment, 7,* 119-129.

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

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

Anderson, P. L. (1982). A preliminary study of syntax in the written expression of learning disabled children. *Journal of Learning Disabilities, 15,* 359-362.

Barenbaum, E., Newcomer, P., & Nodine, B. (1987). Children's ability to write stories as a function of variation in task, age and developmental level. *Learning Disability Quarterly, 7,* 19-29.

Berninger, V. (2000). Development of language by hand and its connections to language

by ear, mouth, and eye. *Topics of Language Disorders, 20,* 65-84.

Berninger, V., & Graham, S. (1998). Language by hand: A synthesis of a decade of research on handwriting. *Handwriting Review, 12,* 11-25.

Berninger, V. W., Vaughan, K. B., Abbot, R. D., Abbot, S. P., Rogan, L. W., Brooks, A., et al. (1997). Treatment of handwriting problems in beginning writers: Transfer from handwriting to composition. *Journal of Educational Psychology, 89,* 652-666.

Berninger, V. W., Vaughan, K. B., Abbot, R. D., Begay, K., Coleman, K. B., Curtin, G., et al. (2002). Teaching spelling and composition alone and together: Implications for the simple view of writing. *Journal of Educational Psychology, 94,* 291-304.

Berninger, V. W., Whitaker, D., Feng, Y., Swanson, H. L., & Abbott, R. D. (1996). Assessment of planning, translating, and revising in junior high writers. *Journal of School Psychology, 34,* 23-52.

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

Bucy, J. E., & Swerdlik, M. E. (1998). Test review of the Test of Written Language, Third Edition. From J. C. Impara & B. S. Plake (Eds.), *The thirteenth mental measurements yearbook* [Electronic version]. Retrieved January 30, 2004, from the Ovid Technologies's *Mental Measurement Yearbook* website: http://web5.silverplatter.com/webspirs/start.ws?customer=c150119&databases=(YB)

Burns, M. K., & Symington, T. (2003). A comparison of the Spontaneous Writing Quotient of the Test of Written Language (3rd ed.) and teacher ratings of writing

progress. *Assessment for Effective Intervention, 28,* 29-34.

Cannon, C. (2006). Implementing research practices. *High School Journal, 89*, 8-12.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18,* 65-81.

Charter, R. A., Walden, D. K., & Padilla, S. P. (2000). Too many simple scoring errors: The Rey Figure as an example. *Journal of Clinical Psychology, 56,* 571-574.

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.

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

Cole, J. C., Muenz, T. A., Ouchi, B. Y., Kaufman, N. L., & Kaufman, A. S. (1997). The impact of the pictorial stimulus on the written expression output. *Psychology in the Schools, 34,* 1-9.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper and L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-31). Urbana, IL: National Council of Teachers of English.

CTB/MacMillian/McGraw-Hill. (1993). *CTB writing assessment.* Monterey, CA: Author.

CTB/McGraw-Hill. (1985). *California Achievement Test.* Monterey, CA: Author.

CTB/McGraw-Hill. (1996). *TerraNova.* Monterey, CA: Author.

Dahl, K. L., & Farnan, N. (1998). *Children's writing: Perspectives from research.* Newark, DE: International Reading Association.

Daiute, C. (1986). Physical and cognitive factors in revising: Insights for studies with computers. *Research in the Teaching of English, 20,* 141-259.

DeGroff, L. C. (1987). The influence of prior knowledge on writing, conferencing, and revising. *Elementary School Journal, 88,* 105-118.

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

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., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. (1982). *The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study.* (Research Report No. 87). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities. Abstract retrieved February 22, 2004, from ERIC database.

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. *Journal of Special Education, 38,* 208-217.

Espin, C., Scierka, B., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading and Writing Quarterly, 14,* 5-27.

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

Faul, F., & Erdfelder, E. (1992). GPOWER (Version 2.0) [Computer program]. Bonn, Germany: Bonn University.

Fewster, S., & Macmillan, D. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education, 23,* 149-156.

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

Gajar, A. (1989). A computer analysis of written language variables and a comparison of compositions written by university students with and without learning disabilities. *Journal of Learning Disabilities, 22,* 125-130.

Gansle, K. A., Noell, G. H., VanDerHeyden, A., 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., Slider, N. J., Hoffpauir, L. D., & Whitmarsh, E. L. (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.

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.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1982). *Stanford Achievement Test-Primary 3*. Austin, TX: Harcourt Brace Javanovitch-The Psychological Corporation.

Good, R. H., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M. R. Shinn (Ed.), *Advances in curriculum-based measurement* (pp. 61-88). New York: Guilford Press.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68,* 532-560.

Graham, S. (1990). The role of production factors in learning disabled students' compositions. *Journal of Educational Psychology, 82,* 781-791.

Graham, S., & Harris, K. R. (1997). It can be taught but it does not develop naturally: Myths and realities in writing instruction. *School Psychology Review, 26,* 414-415.

Grammar slammer. (1997). Retrieved August 8, 2004, from http://englishplus.com/grammar/

Hammill, D. D., & Hresko, W. P. (1994). *Comprehensive Scales of Student Abilities.* Austin, TX: PRO-ED.

Hammill, D. D., & Larsen, S. C. (1978). *The Test of Written Language.* Austin, TX: PRO-ED.

Hammill, D. D., & Larsen, S. C. (1996). *The Test of Written Language – 3rd Edition.*

Austin, TX: PRO-ED.

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

Hansen, J. B. (1997) Test review of the Test of Written Language, Third Edition. From J. C. Impara & B. S. Plake (Eds.), *The thirteenth mental measurements yearbook* [Electronic version]. Retrieved January 30, 2004, from the Ovid Technologies's *Mental Measurement Yearbook* website: http://web5.silverplatter.com/webspirs/start.ws?customer=c150119&databases=(YB)

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test, Ninth Edition.* San Antonio, TX: Author.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinbert (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hidi, S., & McLaren, J. (1990). The effect of topic and theme interestingness on the production of school expositions. In H. Mandl, E. DeCorte, N. Bennett, & H. F. Friedrich (Eds.), *Learning and instruction: European research in an international context* (Vol. 1, pp. 295-308). Oxford, England: Pergamon.

Hidi, S., & McLaren, J. (1991). Motivational factors and writing: The role of topic interestingness. *European Journal of Psychology of Education, 6,* 187-197.

Hooper, S. R. (2002). The language of written language: An introduction to the special issue. *Journal of Learning Disabilities, 35*, 2-6.

Hooper, S. R., Montgomery, J., Swartz, C., Reed, M. S., Sandler, A. D., Levine, M. D., et

al. (1994). Measurements of written language expression. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 375-417). Baltimore: Paul H. Brookes Publishing Co.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996). *Iowa Tests of Basic Skills, Form M.* Itasca, IL: Riverside Publishing.

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.

Hunsely, J. & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15,* 446-455.

Idol, L., Nevin, A., & Paolucci-Whitcomb, P. (1999). *Models of curriculum-based assessment: A blueprint for learning* (3rd ed.). Austin, TX: Pro-Ed.

Isaacson, S. (1988). Assessing the writing product: Qualitative and quantitative measures. *Exceptional Children, 54,* 528-534.

Isaacson, S. (1999). Instructionally relevant writing assessment. *Reading and Writing Quarterly, 14,* 29-48.

Jensen, A. (1980). *Bias in mental testing.* New York: Free Press.

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.

Jones, C. J. (1998). *Curriculum-based assessment: The easy way.* Springfield, IL: Charles C. Thomas.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach.* San Diego: Harcourt Brace Jovanovich.

Kellogg, R. T. (1987). Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory and Cognition, 15,* 255-266.

Lee, L. L., & Canter, S. M. (1971). Developmental sentences scoring. *Journal of Speech and Hearing Disorders, 36,* 335-340.

Li, H., & Hamel, C. M. (2003). Writing issues in college students with learning disabilities: A synthesis of the literature from 1990 to 2000. *Learning Disability Quarterly, 26,* 29-46.

Lloyd, J. W., Weintraub, F. J., & Safer, N. D. (1997). A bridge between research and practice: Building consensus. *Exceptional Children, 63*, 535-538.

MacArthur, C., & Graham, S. (1987). Learning disabled students' composing with three methods: Handwriting, dictation and word processing. *Journal of Special Education*, *21*, 22–42.

Madden, R., Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1978). *Stanford Achievement Test.* New York: Harcourt Brace Jovanovich.

Malecki, C. K., & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools, 40,* 379-390.

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: Guilford Press.

Marston D., Lowry, L., Deno, S., & Mirkin, P. (1981). *An analysis of learning trends in*

simple measure of reading, spelling, and written expression: A longitudinal study. (Research Report No. 49). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities. Abstract retrieved February 22, 2004, from ERIC database.

Marston, D., Mirkin, P., & Deno, S. (1984). Curriculum-based measurement: An alternative to traditional screening, referral, and identification. *Journal of Special Education, 2,* 109-117.

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language, 25,* 431-444.

McCutchen, D., Covill, A., Hoyne, S. H., & Mildes, K. (1994). Individual differences in writing: Implications of translating fluency. *Journal of Educational Psychology, 86,* 256-266.

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

Mehrens, W. A., & Clarizio, H. F. (1993). Curriculum-based measurement: Conceptual and psychometric considerations. *Psychology in the Schools, 30,* 241-254.

Mitzel, H. C., & Borden, C. F. (2000). *LEAP for the 21st century: 1999 operational final technical report.* Monterey, CA: CTB/McGraw-Hill.

Moran, M. R. (1981). Performance of learning disabilities and low achieving secondary students on formal features of a paragraph-writing task. *Learning Disabilities Quarterly, 4,* 271-280.

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

New Jersey Department of Education (2003). *2002 – 2003 New Jersey school report card.* Retrieved April 16, 2004, from http://education.state.nj.us/rc/rc03/dataselect.php?c=23&d=4660&s=055&datasection=all

Parker, R., Tindal, G., & Hasbrouck, J. (1991). Countable indices of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality, 2,* 1-17.

Poplin, M. S., Gray, R., Larsen, S., Banikowski, A., & Mehring, T. (1980). A comparison of written expression abilities in learning disabled and non-learning disabled students in three grade levels. *Learning Disabilities Quarterly, 3,* 46-53.

Poteet, J. A. (1979). Characteristics of written expression of learning disabled and non-learning disabled elementary-school students. *Diagnostique, 4,* 60-74.

Powell-Smith, K. A., & Shinn, M. R. (2004). *Administration and scoring of written expression curriculum-based measurement (WE-CBM) for use in general outcome measurement.* Retrieved July 18, 2004, from http://www.aimsweb.com/uploaded/files/scoring_wecbm.pdf

Reschly, D. (1992). Special education decision making and functional/behavioral assessment. In W. Stainback & S. Stainback (Eds.), *Controversial issues confronting special education: Divergent perspectives* (pp. 286-301). Needham Heights, MA: Allyn and Bacon.

Salvia, J., & Ysseldyke, J. E. (2001). *Assessment* (8th ed.). Boston: Houghton Mifflin.

Shapiro, E. S. (1987). *Behavioral assessment in school psychology.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Shapiro, E. S. (1989). *Academic skills problems: Direct assessment and intervention.*

New York: Guilford Press.

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

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: Guilford Press.

Shinn, M. R., Rosenfield, S., & Knutson, N. (1989). Curriculum-based assessment: A comparison of models. *School Psychology Review, 18,* 299-316.

Spandel, V., & Stiggins, R. J. (1997). *Creating writers: Linking writing assessment and instruction* (2nd ed.). New York: Longman.

Sullivan, K. (2000). Examiners' error on the Wechsler Memory Scale – Revised. *Psychological Reports, 87,* 234-240.

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

Thomas, S. G. (1994). Preparing business students for real-world writing. *Education and Training, 36,* 11-15.

Tindal, G., & Hasbrouck, J. (1991). Analyzing student writing to develop instructional strategies. *Learning Disabilities Research and Practice, 6,* 237-245.

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

Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression.

Learning Disabilities Research and Practice, 6, 211-218.

U.S. Census Bureau (2000). *American FactFinder.* Retrieved April 16, 2004, from http://factfinder.census.gov/servlet/BasicFactsTable?_lang=en&_vt_name=DEC_2000_PL_U_GCTPL_ST7&_geo_id=04000US34

U.S. Department of Education, Institute of Education Sciences, & National Center for Education Statistics. (2003). *The nation's report card: Writing 2000* (NCES 2003-529). Washington, D.C.

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.

Watkinson, J. T., & Lee, S. W. (1992). Curriculum-based measures of written expression for learning-disabled and nondisabled students. *Psychology in the Schools, 29,* 184-191.

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

Wright, J. (1992). *Curriculum-based measurement: Directions for administering and scoring CBM probes in writing.* Retrieved July 18, 2004, from http://www.jimwrightonline.com/pdfdocs/cbmresources/cbmdirections/cbmwrit.pdf

APPENDIX

IRB Approval Letter

PENNSTATE

Vice President for Research
Office for Research Protections

The Pennsylvania State University
212 Kern Graduate Building
University Park, PA 16802-3301

(814) 865-1775
Fax: (814) 863-8699
www.research.psu.edu/orp/

Date: November 24, 2004

From: Tracie L. Kahler, IRB Administrator

To: Janelle M. Amato

Subject: Results of Review of Proposal – Exemption **(IRB #20042) Secondary Data**
Approval Expiration Date: November 23, 2005
"Dissertation"

The Office for Research Protections (ORP) has reviewed and approved your application for the use of human participants in your research. By accepting this decision, you agree to obtain prior approval from the ORP for any changes to your study. Unanticipated participant events that are encountered during the conduct of this research must be reported in a timely fashion.

If your study will extend beyond the above noted approval expiration date, the principal investigator must submit a completed Continuing Progress Report to the ORP to request renewed approval for this research.

On behalf of the ORP and the University, thank you for your efforts to conduct research in compliance with the federal regulations that have been established for the protection of human participants.

TLK/slk
cc: Marley W. Watkins

Curriculum Vita

Janelle (Matesic) Amato
217 Voorhis Avenue
New Milford, NJ 07646
Phone: 201-483-3985
Email: janelleamato@gmail.com

Education:

Doctoral Candidate	School Psychology – anticipated graduation May 2008 Pennsylvania State University, University Park, PA
M.S.	School Psychology - December 2004 Pennsylvania State University, University Park, PA
B.A.	Psychology – May 1999 The College of New Jersey, NJ

Certification:
Certified School Psychologist – Pennsylvania and New Jersey (May 2004)

Work Experience:
School Psychologist, Cresskill School District, Cresskill, NJ (2006 – present)
School Psychologist, Sayreville School District, Sayreville, NJ (2004 – 2006)
Graduate Clinician Supervisor, Pennsylvania State University CEDAR clinic, PA (2003 - 2004)
School Psychology Clinician, Pennsylvania State University CEDAR clinic, PA (2001- 2003)
Data Analyst, PSU Student Affairs Research and Assessment, PA (2001-2004)
Research Consultant, General Assembly of Pennsylvania, PA (2000-2001)
Substitute Teacher, Guttenburg School District, NJ (1999-2000)

Professional Affiliation:
National Association of School Psychologists
New Jersey Association of School Psychologists
American Psychological Association, Division 16 – School Psychology

Publications and Presentations:
Kubina, R. M., Amato, J., Schwilk, C. L., & Therrien, W. J. (in press). Comparing performance standards on the retention of words read correctly per minute. Manuscript accepted for publication to *Journal of Behavioral Education*.

Matesic, J. (2005). In J. T. Neisworth and P. S. Wolfe (Eds.), *The Autism Encyclopedia.* Baltimore: Paul H. Brooks Publishing Company.

Lafer, M., Diamond, N., Matesic, J., Parada, J., Popp, J., & Weaver, J. (2001, November). *Factors affecting time to bachelorette degree at Pennsylvania colleges and universities. Report prepared for the General Assembly of Pennsylvania in response to Senate Resolution 180.* Symposium conducted at the 2001 Association for the Study of Higher Education, Richmond, V.A.

Matesic, J. (2003, March). *The factor structure of the Reynolds Adolescent Depression Scale in adolescents from the Republican of Trinidad and Tobago.* Poster session presented at the 23rd annual spring conferences of the Association of School Psychologists of Pennsylvania, Harrisburg, P.A.