

The Factor Structure of Curriculum-Based Writing Indices
at Grades 3, 7, and 10

by

Alec Judd Brown

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2012 by the
Graduate Supervisory Committee:

Marley Watkins, Co-Chair
Linda Caterino Kulhavy, Co-Chair
Marilyn Thompson

ARIZONA STATE UNIVERSITY

August 2012

ABSTRACT

National assessment data indicate that the large majority of students in America perform below expected proficiency levels in the area of writing. Given the importance of writing skills, this is a significant problem. Curriculum-based measurement, when used for progress monitoring and intervention planning, has been shown to lead to improved academic achievement. However, researchers have not yet been able to establish the validity of curriculum-based measures of writing (CBM-W). This study examined the structural validity of CBM-W using exploratory factor analysis. The participants for this study were 253 third, 154 seventh, and 154 tenth grade students. Each participant completed a 3-minute writing sample in response to a narrative prompt. The writing samples were scored for fifteen different CBM-W indices. Separate analyses were conducted for each grade level to examine differences in the CBM-W construct across grade levels. Due to extreme multicollinearity, principal components analysis rather than common factor analysis was used to examine the structure of writing as measured by CBM-W indices. The overall structure of CBM-W indices was found to remain stable across grade levels. In all cases a three-component solution was supported, with the components being labeled production, accuracy, and sentence complexity. Limitations of the study and implications for progress monitoring with CBM-W are discussed, including the recommendation for a combination of variables that may provide more reliable and valid measurement of the writing construct.

DEDICATION

To my grandfather, Harl Elmer Judd, for leading the way and creating a family culture that values educational achievement and the pursuit of knowledge.

ACKNOWLEDGMENTS

I would like to express gratitude to the people who supported me throughout this process. First, to Dr. Marley Watkins, thank you for continuing to oversee this project even after moving to another institution. Your knowledge and expertise were invaluable, and by holding my work to a high standard you insured that I was successful.

To Dr. Linda Caterino, thank you for always putting your students first and thank you for being willing to step in and provide support when it was needed. To Dr. Marilyn Thompson, thank you for your willingness to assist me even when you were already stretched to your limits with other responsibilities.

Additionally, I would like to express thanks to Dr. David Wodrich and Dr. Amanda Sullivan, who assisted with my proposal, to the principals and teachers in the Chandler Unified School District, who graciously assisted me with data collection, and to the graduate students who assisted in the scoring of writing samples.

To my wife, Jennifer, thank you for the support, encouragement, and time that you devoted to me throughout this project. Your contributions to my success go far beyond what I can express on this page.

TABLE OF CONTENTS

	Page
LIST OF TABLES.....	vi
CHAPTER	
1 INTRODUCTION.....	1
2 REVIEW OF RESEARCH REGARDING CURRICULUM-BASED MEASUREMENT OF WRITING.....	8
Initial Research Regarding CBM-W	10
IRLD Reliability Studies	11
Test-retest Reliability	11
Internal Consistency	11
Alternate Form Reliability	12
Interscorer Agreement	12
IRLD Validity Studies	13
Summary of IRLD Findings.....	14
Subsequent Research on CBM-W.....	15
Subsequent Findings for the Original Scoring Indices	15
Additional CBM-W Indices.....	20
Summary of CBM-W Research.....	27
Other Types of Validity Evidence.....	28
What Structure Might We Expect?	30
Research Questions.....	38
3 METHOD	40

CHAPTER	Page
Participants.....	40
Measures	41
Procedure	42
Data Analysis	45
4 RESULTS	49
5 DISCUSSION	54
Limitations and Future Research.....	58
Conclusion	62
REFERENCES	64
APPENDIX	
A INSTITUTIONAL REVIEW BOARD/HUMAN SUBJECTS	
APPROVAL	114
B CBM-W ADMINISTRATION INTEGRITY SELF-CHECK	116

LIST OF TABLES

Table	Page
1. Descriptions of Indices in Curriculum-Based Measurement of Writing	72
2. Summary of Reliability Studies Conducted at the Institute for Research on Learning Disabilities	76
3. Summary of Validity Studies Conducted at the Institute for Research on Learning Disabilities	78
4. Summary of Additional Studies Examining the Technical Adequacy of Curriculum-Based Measurement of Written Expression	80
5. Sample Demographics by Grade Level	101
6. Interscorer Reliability for CBM-W Indices: Percent Agreement Between Scorers and Primary Investigator	102
7. Means and Standard Deviations for CBM-W Indices at Three Grade Levels	103
8. Third Grade Sample: Correlation Matrix for Curriculum Based Writing Indices	104
9. Seventh Grade Sample: Correlation Matrix for Curriculum Based Writing Indices	106
10. Tenth Grade Sample: Correlation Matrix for Curriculum Based Writing Indices	108
11. Component Loadings for CBM-W Indices for Sample of Third Grade Students: Principal Components Analysis with Promax Rotation ..	110

Table	Page
12. Component Loadings for CBM-W Indices for Sample of Seventh Grade Students: Principal Components Analysis with Promax Rotation.....	111
13. Component Loadings for CBM-W Indices for Sample of Tenth Grade Students: Principal Components Analysis with Promax Rotation ..	112
14. Component Intercorrelations for Principal Components Analysis with Promax Rotation.....	113

Chapter 1

Introduction

In 1983, a now well-known report, *A Nation at Risk: The Imperative for Educational Reform*, highlighted deficits in the American education system and called for corrective reforms (National Commission on Excellence in Education). A number of different accountability and reform movements have subsequently occurred and, unfortunately, the intense focus on reform in some subject areas may have left other areas relatively neglected. For example, the National Commission on Writing (2003) claimed that writing instruction has been neglected in favor of an increased emphasis on mathematics and science instruction in American schools. As a result, most students “cannot write well enough to meet the demands they face in higher education and the emerging work environment” (National Commission on Writing, 2003, p. 16). Results from the National Assessment of Educational Progress (NAEP) seem to support this claim. In 2007, only 33% of eighth grade students and 24% of twelfth grade students scored at or above the proficient level on the writing assessment (National Center for Education Statistics, 2008). Although this represented a slight improvement for 8th grade students when compared to the results of the 2002 NAEP, it is still a dismal result.

For the subset of the population that has learning disabilities, achieving writing competency is even more challenging. In America, 4.2% of students have been identified as having a specific learning disability (Office of Special Education Programs, 2006), a disorder of basic psychological processes that

impairs the ability to read, write, spell, or perform mathematical calculations (Individuals with Disabilities Education Improvement Act, 2004).

Writing difficulties are particularly serious when one considers the importance of writing skills for success in education as well as in the demands of day-to-day life. As stated by Hooper (2002), “writing has become a critical life skill that is intimately linked to basic literacy” (p. 2). In school, writing skills are necessary for the demonstration of knowledge, but even more importantly, writing is a way of thinking through a problem and synthesizing knowledge (Miller, 2009). The National Commission on Writing (2003) has argued that “writing is not simply a way for students to demonstrate what they know. It is a way to help them understand what they know. At its best, writing is learning” (p. 13). Outside of school, writing is a necessary skill for interpersonal communication and for successful functioning in most employment settings. Furthermore, writing competence has significance for society and culture in general, influencing everything from advertisements to movie scripts, and from personal emails and instant messages to poetry (Miller, 2009; National Commission on Writing, 2003).

One reform that has been proposed to address academic concerns for students with learning disabilities and general education students with academic problems is the Response to Intervention (RTI) model. RTI is based on a public health model of service delivery, which focuses on early intervention and prevention through the implementation of evidence-based interventions (Fletcher & Vaughn, 2009). Public health models commonly include three levels of

intervention: (a) primary interventions, which target the entire population, (b) secondary interventions, which target a subset of the population that has been identified as being at risk, and (c) tertiary interventions, which target individuals who have been identified as having the illness or condition in question (Strein, Hoagwood, & Cohn, 2003). Within the framework of RTI, primary interventions are evidence-based teaching methods that are universally applied in all classrooms. Screening measures are used to identify students who are at risk for academic failure and these students receive additional small group instruction, or secondary intervention. Regular progress monitoring is conducted with these students and those who continue to make poor progress receive tertiary interventions that are intensive and individually targeted (Fletcher & Vaughn, 2009; Vaughn & Fuchs, 2003).

To date, there have not been any large-scale studies of the RTI model's effect on writing outcomes; however, the potential benefits of RTI have been demonstrated in other areas. A large-scale study of the RTI model of reading intervention, implemented in 318 high need schools in Florida, demonstrated significant improvement in students' scores on reading assessments and significant reductions in the number of students identified as having learning disabilities (Torgesen, 2009). There is also considerable evidence showing that regular progress monitoring leads to improved academic performance (e.g., Fuchs, Fuchs, & Hamlett, 1989; Stecker, Fuchs, & Fuchs, 2005), and there is mounting evidence that early, intensive intervention can actually normalize the brain activity of children with learning disabilities (e.g., Shaywitz et al., 2004;

Simos et al., 2007). These findings highlight the potential benefits of the RTI model as applied to writing instruction and intervention.

In 2004, the revision of the Individuals with Disabilities Education Act facilitated more widespread implementation of the RTI model by stating that schools “may use a process that determines if the child responds to scientific, research-based intervention as part of the evaluation procedures [for a learning disability]” (Individuals with Disabilities Education Improvement Act, 2004, section 1414(b)(6)). This means that not only can schools use a public health model to provide early intervention, but they can also use this model to help identify students with learning disabilities. Although this change paved the way for more schools to implement RTI, there are many practical considerations that must be addressed before a school can effectively implement the model, not the least of which is, how will schools measure student progress and identify students at risk for academic failure?

Measurement is a key component of the RTI model (Fletcher & Vaughn, 2009; Hosp, Hosp, & Howell, 2007) because there are multiple decision points throughout the RTI process and valid data are needed at each decision point to guide these decisions. Assessment is the basis for (a) identifying through the screening process the students who are at risk, (b) determining if a student is making adequate progress, and (c) making decisions about eligibility for special education (Fletcher & Vaughn, 2009; Vaughn & Fuchs, 2003). Without appropriate measurement tools, there can be no assurance that students are

receiving the level of intervention that they need, or that students with learning disabilities are being accurately identified.

The RTI model relies on curriculum-based measurement (CBM) as the primary assessment tool for screening, progress monitoring, and eligibility decisions (Fletcher & Vaughn, 2009). CBM involves brief assessments of academic skills that can be administered repeatedly over time. CBM assessments are criterion referenced, direct measures of academic skills. These methods are considered well suited for RTI because they are tied closely to the curriculum, are time efficient, and are designed for progress monitoring (Fletcher & Vaughn, 2009; Hosp et al., 2007; Malecki, 2008).

Deno (2003) and Hosp et al. (2007) have specified several important features of appropriate CBM measures. First, these measures must be technically adequate. That is, they should conform to accepted standards for reliability and validity if they are to be used in educational decision-making. Second, the procedures for administering and scoring these measures must be standardized. Third, it must be possible to administer these measures repeatedly and they must be sensitive to change over time because they will be used to monitor progress. Fourth, these measures must be time efficient because they will be administered to large numbers of students on a repeated basis. Fifth, these measures should be aligned with the curriculum. They should also directly sample the behavior of interest, so that it is not necessary to make inference when drawing conclusions about the results. Finally, there should be well-established decision rules for

determining which students are at risk, and whether or not students are making adequate progress.

Curriculum-based measurement of reading (CBM-R) provides a prime example of CBM. The most commonly used CBM-R measures are reading aloud, maze selection, and word identification (Wayman, Wallace, Wiley, Ticha, & Espin, 2007). In reading aloud, the student reads from a passage for one minute and the number of correctly read words is recorded. Word identification is similar, but involves reading aloud from a list of high frequency words. In maze selection, the student reads a passage in which every seventh word has been deleted and replaced with three word choices. The student then selects the word that best fits the context of the passage. An extensive foundation of evidence supports the reliability and criterion-related validity of these measures (Wayman et al., 2007).

Whereas CBM-R measures have proven to be good general outcome measures, or broad measures of skill, curriculum based measures in mathematics (CBM-M) have not (Christ, Scullin, Tolbize, & Jiban, 2008). For example, the most commonly used CBM-M measures are 1- or 2-minute probes that sample basic math facts (Foegen, Jiban, & Deno, 2007). These measures have generally exhibited acceptable reliability, and moderate to strong criterion-related validity when the criterion test primarily measures computation skills; however, they are only weakly correlated with broader measures of mathematics skill (Christ et al., 2008; Foegen et al., 2007). Researchers have introduced other measures in attempts to address this weakness, including problem-solving probes and word

problem probes. Most of these measures have exhibited acceptable reliability and moderate criterion-related validity, but coefficients are not as strong as those for CBM-R, and additional research is needed to “establish a form of CBM-M with greater utility and broader use” (Christ et al., 2008, p. 204).

Although additional research is needed in the areas of CBM-R and CBM-M, Christ et al. (2008) suggested that they are the two most well established CBM procedures. This leaves curriculum-based measurement of writing (CBM-W) as the least established area of CBM. Based on their review of the literature, McMaster and Espin (2007) concluded that extensive research is still needed to identify the most useful procedures for monitoring writing.

Chapter 2

Review of Research Regarding Curriculum-Based Measurement of Writing

Writing is a complex process with many facets and the approach to assessing writing varies depending on the purpose of the assessment and the facet being measured (Hooper et al., 1994). For example, a distinction can be made between direct methods, which require the examinee to produce a writing sample, and indirect methods, which only require the examinee to evaluate certain features of a writing sample (Hooper et al., 1994; Malecki, 2008; Tindal & Parker, 1989a). Another important distinction can be made between subjective scoring procedures, which involve judgment on the part of the rater, and objective procedures, which involve counting quantifiable features of a writing sample (Hooper et al., 1994; Tindal & Parker, 1989a).

CBM-W is a direct measure of written expression that relies on objective scoring procedures. Because CBM-W is a direct measure it aligns well with the type of writing task students will encounter in school and work environments (Hooper et al., 1994; Tindal & Parker, 1989a). Meanwhile, the objective nature of CBM-W allows for greater reliability in scoring than subjective techniques. These features are consistent with the standards for CBM outlined by Deno (2003) and Hosp et al. (2007).

Minor variations exist in CBM-W techniques (e.g., the type of story starter used and the length of time the student is given to write), and researchers continue to examine how these differences impact the reliability and validity of CBM-W scores (e.g., McMaster & Campbell, 2008). However, the following procedure is

generally considered to be best practice for administering CBM-W (Hosp et al., 2007; Malecki, 2008):

1. The examiner provides the student with a written story starter and writing materials.
2. The examiner reads standardized instructions that direct the student to listen to the story starter and then write a story about what happens next.
3. The examiner reads the story starter and gives the student 1 minute to think about what they will write.
4. After 1 minute, the examiner prompts the student to begin writing. The student is given 3 minutes to write, with a reminder at 90 seconds.
5. At the end of 3 minutes, the examiner prompts the student to put down his or her pencil and stop writing.

Once writing samples have been collected, they are scored for quantifiable features such as total words written, words spelled correctly, and so forth (Hosp et al., 2007; Malecki, 2008). A wide variety of these CBM-W indices have been examined in the literature. A list of CBM-W indices and their definitions can be found in Table 1.

CBM-W clearly meets many of the standards for an appropriate curriculum-based measure—it can be administered efficiently and repeatedly, it is aligned with the curriculum, it is a direct assessment, and it has standardized procedures for administration. But does CBM-W have adequate technical adequacy?

This study will use the guidelines for technical adequacy that have been developed by other researchers in CBM (Amato & Watkins, 2011; McMaster & Espin, 2007; Wayman et al., 2007). According to those guidelines, reliability and validity coefficients are considered weak if they fall below .50, moderate if they fall between .50 and .70, and strong if they are .70 or greater.

Initial Research Regarding CBM-W

Curriculum-based techniques for measuring written expression were first introduced by researchers at the Institute for Research on Learning Disabilities (IRLD) at the University of Minnesota in the early 1980's (Deno, Mirkin, & Marston, 1980). They conceptualized CBM-W as part of an instructional methodology that would capitalize on "the stability and generality inherent in repeated assessments of academic skills" (Deno, Marston, & Mirkin, 1982, p. 1) by taking frequent performance samples and tracking a student's progress toward instructional goals to determine the efficacy of teaching strategies. This initial research focused on seven CBM-W indices. The samples for these preliminary studies only included elementary students, and the results were generally promising.

The seven indices included in the IRLD studies were total words written (TWW), words spelled correctly, (WSC), large words (LW), mature words (MW), correct letter sequences (CLS), mean length of T-units (T-units), and correct writing sequences (CWS). Descriptions of each CBM-W scoring procedure are given in Table 1. Some of these indices, such as TWW and WSC, were simple fluency measures, while others were intended to be measures of more complex

writing skill. Long words and mature words, for example, were intended to measure the complexity of vocabulary in a writing sample, while CWS was hypothesized to measure spelling, punctuation, capitalization, and grammar in addition to fluency.

IRLD reliability studies. The primary focus of the IRLD studies was on establishing the reliability and validity of the CBM-W indices. A total of 11 of the IRLD studies examined at least one of the following types of reliability: test-retest, alternate forms, interscorer agreement, or internal consistency. These studies and their results are summarized in Table 2.

Test-retest reliability. Two IRLD studies measured the test-retest reliability of CBM-W indices (Marston & Deno, 1981; Shinn, Ysseldyke, Deno, & Tindal, 1982). The only variable examined in both studies was TWW, which had strong reliability at a 1-day interval ($r = .91$) and moderate reliability at 3 and 4-week intervals ($r_s = .64$ and $.69$). The remaining indices—WSC, CLS and MW—were only examined in the Marston and Deno (1981) study. Both WSC and CLS had strong coefficients at a 1-day interval ($r_s = .81$ and $.92$), but only CLS had good reliability at the 3-week interval ($r = .70$). The 3-week test-retest reliability for WSC was fair ($r = .62$). Reliability coefficients were not acceptable for mature words at either the 1-day or the 3-week interval ($r = .57$ and $.50$).

Internal consistency. Marston and Deno (1981) calculated internal consistency by dividing 5-minute writing samples into 1-minute sections, and then calculating Cronbach's alpha values for CBM-W indices. These values were

acceptable for all four indices examined in their study—TWW, WSC, CLS, and mature words—and ranged from .70 to .87.

Alternate form reliability. Alternate form reliability received greater emphasis than test-retest or internal consistency in the IRLD studies, with a total of five studies examining alternate form reliability for CBM-W indices. Three of these studies found acceptable reliability coefficients ($r_s > .70$) for TWW, WSC, and CLS on comparable story starters (Marston & Deno, 1981; Tindal, Germann, & Deno, 1983; Tindal, Marston, & Deno, 1983). Another study, conducted by Fuchs, Deno, and Marston (1982), took a unique approach to calculating alternate form reliability. The authors administered writing prompts weekly for 10 weeks, and then calculated aggregate alternate forms reliability coefficients (mean WSC for the odd weeks correlated with the mean WSC for the even weeks). Reliability coefficients for WSC improved when aggregated across multiple days, and ranged from moderate when aggregated across 2 days ($r = .55$), to strong when aggregated across 10 days ($r = .89$). This study indicated that the reliability of WSC was greatly improved with multiple samples.

The weakest coefficients were found in a study by Shinn et al. (1982) where four different story starters were administered at 1-week intervals. Reliability coefficients for TWW ranged from .51 to .71. The weaker coefficients found in this study may be accounted for by the fact that there was also a time delay of 1 to 3 weeks.

Interscorer agreement. Research conducted at IRLD consistently found strong inter-scorer reliability coefficients for TWW, WSC, and CLS. In four

different studies, interscorer reliabilities for these three indices ranged from .90 to .99 (Deno et al., 1982; Marston & Deno, 1981; Marston, Deno, & Tindal, 1983; Tindal et al., 1983). The Marston and Deno (1981) study also indicated strong interscorer reliability for mature words ($r = .92$).

Only one of the IRLD studies examined the reliability of CWS (Videen, Deno, & Marston, 1982). This study examined the interscorer agreement for 20 written expression samples scored by two raters. Each individual writing sequence was compared and it was found that the two scorers had an overall percentage agreement of 90.3% for this sample.

IRLD validity studies. IRLD studies examining the validity of CBM-W procedures focused primarily on one type of validity evidence, criterion validity. A variety of criterion measures were used, including the Test of Written Language (TOWL; Hammill & Larsen, 1978), the Stanford Achievement Test (SAT; Madden, Gardner, Rudman, Karlsen, & Merwin, 1978), the Developmental Sentence Scoring System (DSS; Lee & Canter, 1971), and holistic ratings. The results of the IRLD validity studies are summarized in Table 3.

The first effort to establish the validity of CBM-W was a series of three small studies conducted by Deno et al. (1980). The first study used the TOWL as the criterion measure. The resulting mean correlations between the CBM-W indices and the Written Language Quotient of the TOWL were strongest for TWW ($r = .70$) and WSC ($r = .77$), but reasonably strong correlations were also found for MW ($r = .67$) and LW ($r = .62$). Only mean length of T-units had poor criterion-related validity, with a mean correlation of .13. The second study used

the TOWL and the language section of the SAT as criterion measures. Once again, all indices except T-units had strong correlations with the criterion variables. Excluding T-units, correlations with the TOWL ranged from .69 to .83 and correlations with the SAT ranged from .51 to .76. The criterion measure for the third study was the DSS. Results were consistent with the earlier studies. T-units ($r = .29$) and LW were moderately correlated with the DSS ($r = .47$), but the remaining indices were strongly correlated with the DSS.

Researchers at IRLD also examined the criterion-related validity of CWS. Videen et al. (1982) administered a variety of criterion measures to each student, including the TOWL, the DSS, and holistic ratings of the quality of the writing samples (samples were scored by two raters on a scale of 1 to 7). CWS was correlated most strongly with the holistic ratings ($r = .85$). The correlation with the TOWL ($r = .69$) was moderately strong, and the correlation with the DSS was weak ($r = .49$).

Summary of IRLD findings. In summary, the IRLD research demonstrated that T-units had poor criterion-related validity and mature words lacked acceptable reliability; however, the findings for the remaining scoring indices—TWW, WSC, CLS, and CWS—were encouraging. The IRLD reports indicated that these indices generally met standards for acceptable reliability. These indices also appeared to have moderate to strong correlations with a variety of outcome measures. However, it is noteworthy that many of these foundational studies had small samples and none of these foundational studies involved

secondary level students. These factors may explain the difference between IRLD findings and the results of subsequent research.

Subsequent Research on CBM-W

CBM-W research that has been conducted subsequent to the IRLD studies can be categorized in several ways. First, a distinction can be made between studies that have reexamined or extended research on existing scoring procedures versus studies that have introduced new scoring procedures. A second distinction can be made between studies conducted with elementary students versus studies conducted with secondary students. McMaster and Espin (2007) explained the importance of making this second distinction when they stated that many of the simple CBM-W scoring procedures lack sufficient technical adequacy with secondary students and suggested that different scoring procedures may be needed at different grade levels. Table 4 summarizes the studies that have been conducted subsequent to the IRLD studies and have examined the technical adequacy of CBM-W.

Subsequent findings for the original scoring indices. Because TWW, WSC, and CWS were the indices with the strongest support in the IRLD studies, they have been a major focus of subsequent research. The test-retest reliability and internal consistency of these indices have received little attention since the original IRLD studies, perhaps because researchers felt that the reliability of these measures had already been substantiated. The few studies that have examined test-retest reliability and internal consistency have produced positive results. For example, Parker, Tindal, and Hasbrouck (1991b) examined the internal

consistency of TWW, WSC, and CWS in a small sample ($N = 36$) of secondary students and found coefficients ranging from .75 to .78. Gansle, VanDerHeyden, Noell, Resetar, and Williams (2006), in a much larger sample ($N = 538$) of elementary students, obtained test-retest reliability coefficients ranging from .78 to .82 for TWW, WSC, and CWS. The test-retest interval in this study was one week.

Subsequent studies have also found acceptable interscorer agreement for TWW, WSC, and CWS. Watkinson and Lee (1992) found interscorer reliability coefficients that ranged from .95 to .99 for a sample of secondary students, and several other studies obtained very similar coefficients ($r = .86-.99$) for samples of elementary students (Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002; Tindal & Parker, 1991). Gansle et al. (2006) calculated percentage of agreement instead of a reliability coefficient, but the results were similar. Total percentage of agreement for the aforementioned indices ranged from 93.5% to 97.7%.

The results for alternate form reliability have been mixed. For example, Espin et al. (2000) found alternate form reliability coefficients ranging from .72 to .80 for TWW, WSC, and CWS. In contrast, Gansle et al. (2002) found moderate to weak coefficients for the same indices ($r_s = .46-.62$). Two other studies have provided a potential explanation for these inconsistent findings. These studies examined alternate form reliability across grade levels, and found that reliability coefficients were generally acceptable for TWW, WSC, and CWS at the elementary level, but coefficients were weaker at the secondary level, especially for the simple scoring procedures like TWW and WSC (McMaster & Campbell,

2008; Weissenburger & Espin, 2005). The general conclusion was that for the secondary level “more complex scoring procedures applied to longer samples were needed to yield consistently sufficient alternate-form reliability” (McMaster & Campbell, 2008, p. 557).

Whereas subsequent research regarding the reliability of CBM-W has generally been consistent with the IRLD studies, investigations of the validity of CBM-W have not confirmed earlier findings. For example, the original IRLD studies indicated moderate to strong criterion-related validity for TWW and WSC, but subsequent research has produced conflicting results, with validity coefficients typically being weak or non-significant. The strongest coefficients have been found in cases where the criterion measure was holistic ratings of writing quality. In those studies most coefficients have ranged between .35 and .50 (Espin, Scierka, Skare, & Halverson, 1999; Espin et al., 2000; Parker, Tindal, & Hasbrouck, 1991a; Parker et al., 1991b; Tindal & Parker, 1989a, 1989b). Similar coefficients were found when the criterion measure was a district writing assessment ($r_s = .43-.51$; Espin et al., 2000). Weak, but significant, coefficients were also found for language arts and English grades ($r_s = .22-.34$; Espin et al., 1999; Fewster & MacMillan, 2002). Correlations with standardized tests have been lower. TWW and WSC were not significantly correlated with the Woodcock-Johnson-Revised Writing Samples subtest (Gansle et al., 2004; Woodcock & Johnson, 1989), the TOWL (Parker et al., 1991b), or the language section of the Iowa Tests of Basic Skills (Gansle et al., 2002; Hoover, Hieronymus, Fisbie, & Dunbar, 1996), and correlations with the Stanford

Achievement Test, Ninth Edition (Harcourt Brace Educational Measurement, 1996) were weak (Gansle et al., 2006) or non-significant (Jewell & Malecki, 2005).

The findings for CWS have been somewhat more promising, and have consistently produced stronger validity coefficients than TWW and WSC. However, this does not mean that CWS has demonstrated strong criterion-related validity. Once again, the strongest validity coefficients have been found when the criterion measure has been holistic ratings of writing quality. These coefficients have typically been moderately strong at both the elementary level ($r_s = .29-.63$; Parker et al., 1991a; Tindal & Parker, 1991) and the secondary level ($r_s = .45-.83$; Espin, De La Paz, Scierka, & Roelofs, 2005; Espin et al., 1999; Espin et al., 2000; Parker et al., 1991a; Tindal & Parker, 1989a). Weak, but significant, validity coefficients have been found when the criterion measure has been standardized achievement tests, such as the Iowa Test of Basic Skills ($r = .43$; Gansle et al., 2002), the Stanford Achievement Test ($r_s = .41-.43$; Gansle et al., 2006; Tindal & Parker, 1991), the California Achievement Test ($r = .29$; Espin et al., 1999), and the Woodcock-Johnson Writing Samples subtest ($r = .36$; Gansle et al., 2004). One possible explanation for the pattern of higher coefficients for holistic ratings is that they are a direct measure of writing, while most standardized tests are indirect measures. Because CBM-W is a direct measure, we would expect higher correlations with other direct measures.

The studies examined so far have indicated stronger validity for CWS than for the simple production scores (TWW and WSC). Another important

consideration is whether the validity of these measures varies by grade level?

Three studies have examined the validity of CBM-W across grade levels. These studies have used a variety of criterion measures, but in each case the findings have been similar. These studies have all indicated that the validity coefficients for TWW, WSC, and CWS tend to decrease in magnitude as grade level increases, and at every level the validity coefficients for CWS have been stronger than the coefficients for TWW and WSC. At the elementary level all three of these CBM-W indices were significantly correlated with state achievement tests, language arts grades, the Stanford Achievement Test, and analytic ratings (Jewell & Malecki, 2005; McMaster & Campbell, 2008; Weissenburger & Espin, 2005). However, at the secondary level none of the coefficients for TWW and WSC were significant. In comparison, CWS was significantly correlated with the Wisconsin Knowledge and Concepts Examination ($r_s = .47-.52$; Weissenburger & Espin, 2005), which is derived from the TerraNova Assessment Series and the CTB Writing Assessment System (CTB/McGraw-Hill, 1996; CTB MacMillan/McGraw-Hill, 1993). CWS was also significantly correlated with analytic ratings of writing quality ($r = .46$; Jewell & Malecki, 2005) at the junior high level, but correlations with the Test of Written Language, the Stanford Achievement Test, and language arts grades were non-significant (Jewell & Malecki, 2005; McMaster & Campbell, 2008). Only one of these studies included high school students, and it found no significant correlations between either TWW or CWS and the Wisconsin Knowledge and Concepts Examination (WKCE; Weissenburger & Espin, 2005). In summary, studies that have

compared CBM-W indices across grade levels have indicated that TWW and WSC were not valid at the secondary level, and these same studies provided only inconsistent evidence for the validity of CWS at the secondary level. Most importantly, these studies suggested that the criterion-related validity of the standard CBM-W indices decreased as students became older.

Additional CBM-W indices. In light of the disappointing findings regarding the validity of the original CBM-W indices, researchers have explored a variety of alternative scoring procedures. Tindal and Parker (1989a) were the first to examine alternative CBM-W indices. They correlated eight CBM-W indices with four judges' mean holistic ratings of the same writing sample. The sample for this study included 172 students in grades six through eight. The indices used in this study included three of the original scoring procedures--TWW, WSC, and CWS--and the following additional indices: legible words (LegW), mean length of correct writing sequences (ML/CWS), percentage of words spelled correctly (%WSC), percentage of correct writing sequences (%CWS), and percentage of legible words (%LegW). Definitions of these variables are given in Table 1.

LegW showed little promise in Tindal and Parker's (1989a) study. Correlations between LegW and the holistic ratings were significant, but weak ($r = .24$). Only one other study has examined the criterion-related validity of LegW. The study was conducted with a small sample of junior high students ($N = 36$). Correlations with the TOWL were not significant and correlations with holistic ratings were once again weak, although larger than the previous study ($r = .45$; Parker et al., 1991b).

The next scoring procedure, ML/CWS, showed more promise, with a moderately strong correlation between ML/CWS and holistic ratings ($r = .59$; Tindal & Parker, 1989a). As a result, three other studies have included ML/CWS in their analyses. Parker et al. (1991b) also found a moderately strong correlation with holistic ratings ($r = .63$), but correlations with the TOWL were not significant. Espin et al. (1999), who conducted a study with 147 high school students in remedial programs, found that ML/CWS was weakly correlated with the language section of the California Achievement Test ($r = .34$; CAT; CTB/McGraw-Hill, 1985) and holistic ratings ($r = .40$), but was not significantly correlated with English grades. However, the most important findings were related to reliability. Parker et al. (1991b) found poor test-retest reliability ($r_s = .26-.66$), and Espin et al. (2000) found that alternate form reliability was so poor ($r_s = .32-.57$) that they chose to exclude ML/CWS from their validity analyses altogether.

The final three scoring procedures examined by Tindal and Parker (1989a) were CBM-W indices that had been converted to ratios. The results of the Tindal and Parker study indicated a weak correlation between %LegW and the holistic ratings ($r = .42$), but the correlations between the other two percentage measures and the holistic ratings of writing quality were strong (%WSC, $r = .73$; %CWS, $r = .75$). Because the percentage scores had such strong validity coefficients in this initial study, researchers have continued to examine their utility.

In general, the percentage indices have demonstrated acceptable reliability. Several studies have indicated acceptable interscorer reliability

(Parker et al., 1991b; Tindal & Parker, 1989a; Watkinson & Lee, 1992), and internal consistency (Parker et al., 1991b). These indices have also exhibited acceptable test-retest reliability when the test interval was one month ($r_s = .75-.76$), but weaker coefficients have been found for longer intervals ($r_s = .17-.46$; Parker et al., 1991b).

Excluding Tindal and Parker's study (1989a), criterion-related validity coefficients have been moderately strong. Parker et al. (1991b) found moderately strong correlations between %LegW and the TOWL in a small sample of junior high students with learning disabilities ($r = .56, N = 36$). They also found that both %LegW and %WSC were moderately correlated with holistic ratings ($r_s = .53-.60$). The same researchers found similar results in a much larger sample of general education students ($N = 2,160$; Parker et al., 1991a). Using holistic ratings as the criterion, coefficients for %CWS and %WSC were moderate to strong in their elementary sample ($r_s = .43-.70$). Meanwhile, coefficients for junior high and high school students were weak ($r_s = .34-.46$). Jewell and Malecki (2005) found that validity coefficients decreased in strength for higher grades. They found weak to moderate correlations with SAT language scores ($r_s = .46-.67$), language arts grades ($r_s = .29-.58$), and analytic ratings of writing quality ($r_s = .34-.49$), with coefficients that were consistently lower for junior high students as opposed to elementary students. Amato and Watkins (2011) found moderately strong correlations between %CWS and the TOWL Writing Quotient ($r = .61$) in an eighth grade sample. They also found that of the 10 CBM-W indices included in their study, %CWS contributed the most unique

variance to the prediction of TOWL scores. In their sample %WSC was weakly correlated with the TOWL ($r = .41$).

Overall, the percentage indices have shown some promise, but they also have weaknesses. First, as the studies discussed above suggest, these indices show the same pattern of weaker validity coefficients at higher grade levels that has been seen with other indices. Second, their utility for progress monitoring is questionable. For example, Malecki and Jewell (2003) found that %CWS and %WSC scores were not significantly different for elementary versus junior high students. These results indicate that the percentage indices may not distinguish between students at different levels, and may not be sensitive to growth.

McMaster and Espin (2007) explained that this might be due to the following characteristics:

percentage measures do not have equal interval scales and are thus difficult to interpret when trying to distinguish among students at different skill levels. Moreover, they are problematic for monitoring progress (e.g., if a student produced 10 WSC out of 20 WW in fall, and 50 WSC out of 100 WW in spring, %WSC would not reflect any growth, possibly masking important progress). (p. 79)

Two other groups of researchers that introduced alternative CBM-W indices were Espin et al. and Gansle et al. The variables introduced by Espin et al. were characters, characters per word, sentences, words per sentence, and correct minus incorrect writing sequences (Espin et al., 1999; Espin et al., 2000). Most of these variables have only received attention in Espin's studies, perhaps

because they showed little promise. For example, characters per word was found to have unacceptably low alternate form reliability ($r_s = .12-.47$), and although characters had acceptable reliability, it had weak criterion-related validity (Espin et al., 1999; Espin et al., 2000). Words per sentence had a strong negative correlation with a district writing test ($r_s = -.61$ to $-.76$), but correlations with holistic ratings were weak ($r_s = -.39$ to $.37$; Espin et al., 2000). Sentences showed more promise, with strong validity coefficients when a district writing test was the criterion variable (Espin et al., 2000) and moderately strong coefficients when holistic ratings were the criterion variable (Espin et al., 1999; Espin et al., 2000). Correlations with the CAT and English GPA were significant but weak (Espin et al., 1999).

Of the variables introduced by Espin and colleagues, the most extensively studied has been correct minus incorrect writing sequences (CIWS). This variable is a variation on CWS that also accounts for errors, resulting in a measure of writing accuracy. The first published study to examine CIWS involved a sample of junior high students ($N = 112$; Espin et al., 2000). The study found moderate to strong validity coefficients for both criterion variables—holistic ratings ($r_s = .65-.70$) and a district writing test ($r_s = .69-.75$). These results were particularly encouraging considering the fact that most CMB-W indices have shown weak or non-significant correlations with criterion variables at the secondary level.

Subsequent studies have provided additional support for the validity of CIWS. At the elementary level CIWS has been moderately to strongly correlated with holistic ratings and teacher rankings ($r_s = .43-.84$; Lembke, Deno, & Hall,

2003), state writing tests ($r_s = .54-.68$, McMaster & Campbell, 2008; Weissenburger & Espin, 2005), language arts grades ($r = .61$; Jewell & Malecki, 2005) and the SAT language subtest ($r_s = .57-.62$; Jewell & Malecki, 2005). At the secondary level CIWS has shown strong correlations with holistic ratings of writing quality ($r_s = .67-.82$), but the sample size was small ($N = 22$; Espin et al., 2005). Correlations with state writing assessments have been moderately strong at the junior high level ($r_s = .60-.63$) and weak at the high school level ($r_s = .29-.36$; Weissenberger & Espin, 2005). A moderately strong correlation was found between CIWS and the TOWL Writing Quotient ($r = .56$; Amato & Watkins, 2011). Weak correlations were also found with the SAT language subtest ($r = .41$) and language arts grades ($r = .36$) in a junior high sample (Jewell & Malecki, 2005). Once again, studies conducted across grade levels have found the same pattern that has been present with other indices, namely that validity coefficients decrease in magnitude as grade level increases (Jewell & Malecki, 2005; McMaster & Campbell, 2008; Weissenburger & Espin, 2005). However, this pattern is not as pronounced for CIWS as it is for simple production indices such as TWW and WSC.

Gansle and colleagues have also introduced a variety of alternative scoring procedures. In 2002, Gansle et al. conducted an exploratory study of a large number of new CBM-W indices, including parts of speech, long words, total punctuation marks (TPM), correct punctuation marks (CPM), correct capitalization (CC), complete sentences (CS), words in complete sentences (W/CS), sentence fragments (SF), and simple sentences (SS). Definitions of these

variables are given in Table 1. The sample for this study was composed of 179 third and fourth grade students, and the criterion variables were the Iowa Test of Basic Skills (ITBS; Hoover et al., 1996), the Louisiana Educational Assessment Program (LEAP; Mitzel & Borden, 2000), and teacher rankings of writing proficiency. The results indicated that the majority of these measures lacked sufficient technical adequacy. Of these measures, only CS ($r = .62$) and CPM ($r = .59$) had alternate form reliability coefficients above .50. Furthermore, only four of these variables—CPM, W/CS, SS, and TPM—were significantly correlated with more than one of the criterion variables.

Despite these lackluster results, several of these variables have received further examination, but only the variables related to punctuation and complete sentences have shown promise. Test-retest reliability coefficients for CPM, CS, and W/CS have approached the standard for acceptable reliability ($r_s = .61-.65$; Gansle et al., 2006) and CPM's alternate form reliability has been shown to be acceptable in a high school sample ($r = .76$; Diercks-Gransee, Weissenburger, Johnson, & Christensen, 2009). In regards to validity, CS and W/CS were significantly correlated with the SAT language subtest ($r_s = .36-.41$; Gansle et al., 2006), but were not significantly correlated with the Woodcock-Johnson Writing Samples subtest (Gansle et al., 2004). On the other hand, the punctuation measures have been shown to be significantly correlated with both the SAT, the Woodcock-Johnson Writing Samples subtest (Gansle et al., 2004; Gansle et al., 2006) and the TOWL Writing Quotient (Amato & Watkins, 2011). Furthermore, in a sample of high school students CPM had a moderately strong and significant

correlation with holistic ratings ($r = .62$), and a significant, albeit weak, correlation with the language arts portion of the WKCE ($r = .28$; Diercks-Gransee et al., 2009). This last finding is particularly encouraging given the pattern of decreasing validity coefficients at higher grade levels for most CBM-W indices.

Several other index scores that have been examined are adverbs, adjectives, and incorrect writing sequences (IWS). Adverbs and adjectives have only been examined in one study, in which they had extremely low alternate form reliability and non-significant correlations with both holistic ratings and the WKCE (Diercks-Gransee et al., 2009). On the other hand, in a high school sample IWS demonstrated acceptable alternate form reliability, moderately strong negative correlations with the WKCE, and strong correlations with holistic ratings (Diercks-Gransee et al., 2009).

Summary of CBM-W research. In summary, researchers' efforts to find reliable and valid CBM-W indices have only been partially successful. Many of these indices have demonstrated acceptable reliability. In regards to validity, the initial IRLD studies were promising, but when the entirety of CBM-W research is considered, support for the validity of CBM-W is only moderately strong. Among the most extensively studied scoring procedures, those that measure simple production of text, such as TWW and WSC, appear to have some utility at the elementary level, but lack the necessary technical adequacy at the secondary level (McMaster & Espin, 2007). Scoring procedures that are more complex, such as CWS and CIWS, appear to be more appropriate for use at the secondary level than simple production measures, but most studies indicate that these indices

have only weak to moderate criterion-related validity. Percentage measures also appear to have moderately strong validity evidence, but they may not be appropriate for progress monitoring. A number of other variables, such as CPM and W/CS, have some preliminary evidence for their reliability and validity, but need additional research to substantiate their utility. Overarching all of these findings is a general pattern of decreasing magnitude of validity coefficients as grade level increases. This pattern has been present for all variables in all studies examining the technical adequacy of CBM-W indices across grade levels. These results seem to suggest that the writing process becomes more complex as students mature, likely necessitating the use of measures that are more complex, or the use of a combination of measures at higher grade levels (McMaster & Espin, 2007).

Other Types of Validity Evidence

As indicated by the preceding review, the vast majority of CBM-W research has involved examinations of reliability and criterion-related, or external validity, and although establishing the criterion-related validity of CBM-W is critical, it constitutes only one aspect of validity. According to Messick (1995), validity is a unified concept that involves six aspects of validity evidence. Establishing the validity of an assessment instrument involves compiling empirical evidence for various aspects of validity and then making a rational argument for the test's specific use based on the evidence. Therefore, the strongest case for the validity of any assessment method is made when multiple types of validity evidence are gathered and integrated.

One type of validity evidence that has received very little attention in the CBM-W literature is structural validity. The structural aspect of validity refers to how well the internal structure of an assessment represents the structure of the targeted construct (Messick, 1995). Writing is a complex process that entails a variety of tasks, and relies on a number of different cognitive processes (Berninger, Whitaker, Feng, Swanson, & Abbott, 1996; Hayes, 1996; Hayes & Flower, 1980). As such, it is important for writing assessments, such as CBM-W, to measure a variety of tasks and processes; otherwise the validity of the assessment will be threatened by construct underrepresentation (Messick, 1995).

A common method for evaluating the structural aspect of validity is factor analysis. This technique has already been applied to other areas of CBM. Thurber, Shinn, and Smolkowski (2002) examined the factor structure of CBM math measures (CBM-M) for a sample of 207 fourth grade students. Their purpose was to determine whether CBM-M functioned as a general measure of math achievement, or whether it was primarily a measure of computation or applications. Thurber et al. (2002) felt that their research questions and existing math theory were sufficient to allow them to specify several factor models. Consequently, they used confirmatory factor analysis to compare the fit of competing models. The results indicated that the most defensible model was a two-factor model where CBM-M loaded on the computation factor, not the application factor.

A similar study was conducted to determine the aspect of reading to which CBM-R was most strongly related (Shinn, Good, Knutson, & Tilly; 1992). As in

the CBM-M study, confirmatory factor analysis was used because extensive theory regarding the structure of reading already existed and the research questions were specific enough to develop several competing models. The sample for this study included 114 third grade students and 124 fifth grade students. Separate analyses were conducted for each grade level. For the third grade sample a single-factor model, where CBM-R loaded on the general reading factor, was found to have the best fit; whereas a two-factor model, with CBM-R loading on the decoding factor was found to have the best fit for the fifth grade sample.

These studies illustrate how examining the factor structure of a measure can provide valuable information. In the case of CBM-W, examining structural validity may help determine whether the modest criterion-related validity coefficients are the result of construct underrepresentation. An examination of structural validity may also provide insight into the pattern of decreasing validity coefficients across grade levels. Specifically, it may indicate whether the writing construct increases in complexity as grade level increases.

What structure might we expect?

Evaluating the structural aspect of validity involves comparing the statistical structure of the assessment tool to the structure that is expected based on our knowledge of the construct of interest (Messick, 1995). What, then, is the structure of the writing process?

The most influential model of the writing process was introduced by Hayes and Flower (1980). They proposed that writing involves three cognitive

processes—planning, translating, and revising—all of which operate within the context of the task environment and engage the individual’s long-term memory. Hayes (1996) later revised this model to take into account subsequent empirical evidence. His updated theory still conceptualized writing as an interactive process between the individual and the task environment, but several important revisions were made. First, the individual’s motivation and affect were added to the model as important factors influencing the writing process. Second, several changes were made to the three cognitive processes involved in writing. Revision was replaced with text interpretation, while planning and translating were subsumed under broader categories labeled reflection and text production.

- *Reflection* can be generally described as the process of generating and organizing ideas. It involves planning, problem solving, decision-making, and inferencing.
- *Text production* involves retrieving semantic content from long-term memory, forming portions of sentences in working memory, and then transcribing those sentences into writing.
- *Text interpretation* involves reading and evaluating what has been written and then revising as necessary. Text can be evaluated and revised on either a local level (problems at the sentence level, such as conventions and grammar) or a global level (e.g. organization and flow of ideas).

Finally, and most importantly, working memory was added to the model and acknowledged as having a central role in the writing process. According to

Hayes' model (1996), information continually flows between long-term and short-term memory as a person engages in the various writing processes.

Hayes' model provides a theory regarding the cognitive processes involved in writing, but should we expect any CBM-W index to tap into these distinct processes? In order to accomplish its purposes, CBM-W must be brief to administer, but this may limit the scope of the assessment. For example, it is not likely that students will engage in more than minimal revision (text interpretation) when they are given only three minutes to write.

Several other writing theories may also be relevant to this question. Berninger et al. (1997) contended that Hayes' theory underestimates the importance of transcription, particularly in the case of young or unskilled writers. For these writers, the process of transcription (the mechanics of translating thoughts to writing, including handwriting, punctuation, and spelling) may place such a heavy burden on working memory capacity that very few resources remain for reflection (e.g., planning, organizing). Furthermore, Berninger et al. (1996) pointed out that young writers are more likely to use "knowledge-telling" procedures where they simply write down whatever information they are able to recall relevant to the topic, rather than engaging in the reflective processes described by Hayes. Based on Berninger's ideas, we may expect to find that the CBM-W indices have a very simple factor structure at the elementary level, with text generation accounting for the majority of the variance in writing ability.

At higher grade levels, a more complex factor structure would be expected. As students mature and their transcription skills become automatized,

placing less burden on working memory, writers are able to devote more resources to higher level processes such as generating ideas, planning, and revising. Although a more complex factor structure is expected at the secondary level, it is unclear what structure to expect. The structure could align with Hayes' writing theory, but realistically it seems unlikely that short writing samples and simple scoring procedures will measure underlying cognitive processes such as working memory and reflective processes. It may be more likely that CBM-W will simply measure different components of a writing sample, rather than measuring the cognitive processes involved in producing the sample. For example, Bradley-Johnson and Lesiak (1989) proposed five components of writing that they felt were important to the assessment of writing. The components they identified were mechanics (i.e., handwriting), production, conventions, linguistics, and cognition (i.e., organization). It is possible that a factor structure of CBM-W indices will cluster in a manner consistent with these elements.

Several exploratory factor analyses of CBM-W indices have been completed and although they each had limitations, they also provide some indication of the factor structure that may be expected. Three factor analyses have been conducted with elementary populations. Tindal and Parker (1991) conducted a principal components analysis with Varimax rotation on nine variables, six of which were CBM-W indices (the other three variables were analytic ratings of certain elements of the writing sample). Their sample included 211 regular and special education students in grades 3 through 5. Their results

indicated a three-factor solution that included a simple *production factor*, an *accuracy factor*, and a factor measuring the *quality of ideas*. Consistent with Berninger's theory, the production factor accounted for the largest portion of the variance in this elementary sample. Of the six CBM-W indices, four loaded on the production factor (TWW, WSC, CWS, and total word sequences) and two loaded on the accuracy factor (incorrect writing sequences and %CWS). None of the CBM-W indices loaded on the third factor.

The second exploratory factor analysis with an elementary sample was conducted by Puranik, Lobardino, and Altmann (2008). Their sample included 120 students in grades 3 through 6. This study used written retell, rather than a narrative story starter, and the majority of the variables were not common CBM-W variables, which may limit the study's application to CBM-W. However, the results still provide some insight into the expected structure of direct writing measures. The variables included in the study were TWW, total number of ideas expressed, T-Units, mean length of T-Units, number of clauses, clause density, and percentage of grammatically correct T-Units (of these variables only TWW and mean length of T-Units have been used for CBM-W). A principal components analysis with Oblimin rotation was used to examine the factor structure of the writing variables. Results indicated a three-factor solution that was very similar to Tindal and Parker's (1991) findings. Puranik et al. labeled the components *productivity*, *accuracy*, and *complexity*. Once again, the productivity factor accounted for the largest portion of the variance.

Another relevant factor analysis was recently completed by Wagner et al. (2011). Their study used confirmatory factor analysis, rather than exploratory factor analysis, to further examine the factor structure indicated in the study by Puranik et al. (2008). Participants for the study were 208 first and fourth grade students. Participants completed compositional writing samples that were scored for 10 writing variables, including three macro-organizational variables that were scored using subjective ratings and seven countable indices. The countable indices included mean length of T-Unit, clause density, TWW, number of different words, and number of spelling and capitalization errors. Three models were compared—a general model, a two-factor model (macro level and micro level), and a four-factor model. The four-factor model included the three factors found by Puranik et al. (2008), productivity, accuracy and complexity, and a fourth, macro-organization factor. At both grade levels the four-factor model had substantially and significantly better fit to the data than the other two models. Although this study only included two CBM-W indices, it provided further indication of the potential factor structure of direct writing measures.

Only one exploratory factor analysis has been conducted with a secondary sample. The sample for the study was composed of 172 sixth through eighth grade students receiving remedial or special education (Tindal & Parker, 1989a). Eight CBM-W variables were included in the study and the factor structure was examined using common factor analysis with Varimax rotation. Unlike the other studies, this study produced a two-factor solution. Four variables loaded on each factor, with TWW, WSC, CWS and LegW loading on the first factor and %CWS,

%WSC, %LegW, and ML/CWS loading on the second factor. Tindal and Parker labeled these factors *production dependent* and *production independent*. As the name implies, the production dependent variables seem to measure text generation, while the production independent variables seem to measure accuracy. Therefore, the results seem to align well with the results of their factor analysis with the elementary sample (Tindal & Parker, 1991). The results may seem inconsistent with the expectation of a more complex factor structure for secondary students; however, it is important to remember that this sample only included struggling writers who were still at a more basic level of writing proficiency.

Taken together these results give several indications regarding the factor structure that may be expected in a comprehensive examination of the factor structure of CBM-W indices. First, they indicate that a production factor may indeed be an important factor at the elementary level, explaining a large portion of variance, as Berninger's theory would suggest. Second, the fact that two and three-factor solutions were indicated supports the possibility that direct, objective measures of writing may be able to tap into various aspects of the complex writing process.

These results also highlight several gaps in previous studies. First, no study to date has compared the factor structure of CBM-W indices across grade levels. Second, none of these studies has included a comprehensive set of CBM-W variables. The analyses conducted by Tindal and Parker (1989a, 1991) included a limited number of variables, and predated the introduction of several promising variables including CIWS, CPM, and W/CS.

Even more problematic are multiple weaknesses in the factor analytic methods employed in these studies. All three studies suffered from many of the common methodological shortcomings identified by Preacher and MacCallum (2003), and Fabrigar, Wegener, MacCallum, and Strahan (1999). First, it can be argued that the two studies that employed Principal Components Analysis (PCA) used the wrong type of analysis given that the purpose for conducting factor analysis was to examine the underlying structure of writing variables. PCA is a data reduction technique that is appropriate when the goal is to obtain a smaller set of composite variables (or components) that explain as much of the original variance as possible. However, when the goal is to explain the correlations between variables in terms of underlying latent constructs, as was the case in these studies, Exploratory Common Factor Analysis (ECFA) should be used (Fabrigar et al., 1999). A second concern is the method that was used to decide how many factors to retain. Puranik et al. (2008) used the Kaiser-Guttman rule and the scree test, but the Kaiser-Guttman rule has been shown to be susceptible to both underestimating and overestimating the number of factors to retain (Preacher & MacCallum, 2003). Tindal and Parker (1989a, 1991), on the other hand, did not describe the procedures used to guide their decision. Another shortcoming of these studies was the rotation method. Tindal and Parker (1989a, 1991) selected Varimax rotation, an orthogonal rotation method, in both of their studies. Orthogonal rotations constrain factors to be uncorrelated, but it is unlikely that the various writing factors would be unrelated to one another. In cases where such a constraint is not theoretically defensible, orthogonal rotation is

“unwarranted and can yield misleading results” (Fabrigar et al., 1999). Finally, these researchers failed to address multicollinearity, a potential problem with several of the variables included in their studies. All three studies included one or more pairs of measures having bivariate correlations greater than .90, which indicate potential problems with multicollinearity (Child, 2006; Field, 2009). Multicollinearity is a significant problem in factor analysis because it can result in unstable factor solutions (Pett, Lackey, & Sullivan, 2003). Future research should address these limitations.

Research Questions

At the present time, efforts to establish the validity of CBM-W have been limited almost entirely to examinations of criterion-related validity. However, establishing the validity of any test should involve a variety of validity evidence, including structural validity (Messick, 1995). An examination of the structure of writing as measured by CBM-W is important because it will help determine whether the CBM-W indices currently in use are able to measure multiple elements of the complex writing process. Identifying the factor structure of CBM-W may also indicate whether a combination of indices is needed to accurately measure writing skill at various grade levels, and if so, it may guide the selection of appropriate indices. Furthermore, by comparing the factor structure across grade levels, a factor analysis may provide insight into the reason for decreasing validity coefficients for higher grade levels. Therefore, the proposed research questions were:

- What is the factor structure of writing as measured by the CBM-W indices commonly used in research and practice?
- Is the factor structure consistent across elementary, middle, and high school levels, or is the structure more complex at higher grade levels?

Based on the preceding literature review, it was hypothesized that the factor structure would be more complex at higher grade levels than at lower grade levels. At the elementary level, where writing is constrained by text generation and translation skills, it was expected that a production factor would account for the majority of the variance in writing skill, whereas at the junior high and high school level it was expected that the role of production would be diminished.

Chapter 3

Method

Participants

Participants for this study were 561 students from grades three ($n = 253$), seven ($n = 154$), and ten ($n = 154$) recruited from the general student population in a suburban, southwestern school district. These grade levels were selected to reflect the pattern of decreasing validity coefficients seen when comparisons are made between elementary, junior high, and high school students (McMaster & Campbell, 2008; Weissenburger & Espin, 2005). The selection of these grade levels was also guided by research on the developmental differences in the writing process, which includes decreased importance of transcription at higher grade levels (Berninger, 1999; McCutchen, 2006), the increased use of higher-order skills such as planning and revising beginning in early adolescence (Berninger et al., 1996; McCutchen, 2006), and the transition from learning-to-write to writing-to-learn that occurs between the early and upper grades (Berninger, Garcia, & Abbott, 2008).

Participants were recruited from four elementary schools, two junior high schools, and two high schools. The total sample was 44% male and 56% female. Sixty-two percent of the participants were Caucasian, 21% were Hispanic, 12% were Asian/Pacific Islander, 5% were African American, and less than 1% were Native American. Nine percent of the participants were classified as receiving special education services. Approximately 1% were classified as English language learners. The mean age of the third grade sample was 9 years 3 months,

the mean age of the seventh grade sample was 13 years 0 months, and the mean age of the tenth grade sample was 16 years 1 month. Table 5 summarizes demographic information for each grade level.

Measures

The variables included in the factor analysis were 15 of the most common CBM-W indices. Three-minute writing samples were collected. Each CBM-W probe was scored for 15 indices: characters, correct letter sequences, total words written, words spelled correctly, complete sentences, correct minus incorrect writing sequences, correct punctuation marks, legible words, mean length of correct writing sequences, percentage of correct writing sequences, percentage of legible words, percentage of words spelled correctly, sentences, words per sentence, and mean length of T-Units. These variables are defined in Table 1. The reliability and validity of these indices has already been discussed.

These indices were selected using several criteria. First, there had to be at least minimal evidence for the reliability and validity of the index. This was defined as at least one study indicating acceptable test-retest or alternate form reliability (i.e., greater than .65), and at least one study indicating criterion-related validity greater than .30. These fairly liberal standards were used in order to insure a comprehensive sampling of CBM-W indices and to obtain sufficient variables to insure that each factor was overdetermined.

Sixteen variables met the criterion of minimal evidence for reliability and validity, but it was necessary to exclude two variables due to singularity. CIWS is a linear combination of CWS and IWS, meaning that the variables are perfectly

correlated. The decision was made to retain CIWS because it has the strongest evidence for validity across grade levels. Accordingly, CWS and IWS were removed from the list of scoring procedures, leaving 14 variables.

The decision was also made to add one index, mean length of T-units, to the analysis, resulting in the final set of 15 variables. Mean length of T-units did not meet the inclusion criteria because its reliability has not been examined in CBM-W studies. Despite this omission, mean length of T-units was included because previous studies found that it loaded on a complexity factor (Puranik et al., 2008; Wagner et al., 2011). As such, it was felt that it might serve as a marker variable for locating other variables in factor space (Pett et al., 2003). Furthermore, although test-retest and alternate form reliability have not been examined, there is evidence that mean length of T-units has acceptable interscorer reliability (Puranik et al., 2008; Wagner et al., 2011).

Procedure

Participants were recruited from the general student population in a southwestern school district after approval was obtained from the school district and the university institutional review board. First, site administrators were contacted to obtain approval to conduct the research at specific school sites. Once the site administrator had provided approval, informed consent letters were sent to the parents of all students in the selected grade level. Only students whose parents provided consent were included in the study. The participants also provided written assent prior to participation.

A single CBM writing probe was administered to each participant. A narrative story starter was used, because it is the most common type of CBM-W prompt. The same prompt was used for all grade levels in order to avoid variability due to differences between story starters: “One day our teacher was sick. We had a substitute teacher and. . . .”

The participating students were brought to a central location, such as the school library, and the probe was group administered during regular school hours. The standard administration procedure for CBM-W, as described by Malecki (2008), was followed. The examiner provided the students with the written story starter and writing materials, and then read the following standardized instructions:

You are going to write a story. First, I will read a sentence, and then you will write a story about what happens next. You will have one minute to think about what you will write, and three minutes to write your story. Remember to do your best work. If you don't know how to spell a word, you should guess. Are there any questions? [Pause] Put your pencil down and listen. For the next minute, think about . . . (p. 478)

Next, the examiner read the story starter; began the stopwatch, and gave students one minute to think. After 30 seconds students were given a reminder to think about the story starter. At the end of one minute, the students were prompted to begin writing. The stopwatch was restarted, and the students were

given three minutes to write. A reminder was given after 90 seconds. At the end of three minutes students were told to stop and put down their pencils.

A self-check form was used to ensure integrity of CBM-W administration. The form listed the steps in the administration procedure described above. After each step, the examiner recorded whether the procedure had been correctly followed. Administration integrity was 99.8%. In one instance the examiner failed to provide the reminder given at 90 seconds.

The writing samples were independently scored by three graduate students—two in school psychology and one in speech-language pathology/audiology. The primary investigator trained the scorers in a session that lasted approximately two and one half hours. At the end of the training session the scorers practiced scoring three CBM probes and were provided with specific feedback on any scoring errors that they made. Once trained, they scored a fourth protocol. Scorers were required to obtain scoring accuracy of 90% or greater on the fourth probe prior to scoring student protocols. Scorers were also provided with a set of instructions describing each scoring procedure, which they used as a reference during scoring.

One in every 10 CBM-W probes was randomly selected and independently scored by the primary investigator to ensure that interscorer reliability remained high. Interscorer reliability was measured by percent agreement with the primary investigator's ratings. If agreement fell below 90%, the packet of 10 probes was rescored. Using these procedures it was necessary for seven packets to be rescored. Before the packets were rescored the primary

investigator provided the scorer with additional training regarding the specific errors that had been made. Once the packet had been rescored, a second writing sample was randomly selected from the packet and percent agreement was calculated again for the variables that had lacked sufficient agreement.

Interscorer reliabilities are reported in Table 6. As indicated in the table, total percent agreement exceeded 90% for all variables except Incorrect Writing Sequences (IWS) and strings of correct writing sequences. Interscorer reliability was 80.1% for IWS and 84.2% for strings of correct writing sequences. The lower agreement rate for IWS can be attributed to the fact that many writing samples contained a small number of errors, and in those cases a single disagreement would lead to a low percent agreement for IWS. Despite the lower percent agreement for IWS, the percent agreement for Correct Minus Incorrect Writing Sequences (CIWS) was high (94.3%). A single disagreement on IWS could also lead to a different count for strings of correct writing sequences, and once again since the number of strings was generally low a single difference would lead to a low percent agreement.

Data Analysis

Because there is not a clear theory regarding the expected factor structure of CBM-W, and because previous exploratory factor analyses could not be relied on to guide model construction due to methodological flaws; exploratory, rather than confirmatory, factor analysis was considered the most appropriate method for examining the factor structure of CBM-W (Fabrigar et al., 1999). The

structure of the CBM-W indices was examined by conducting three separate factor analyses, one at each grade level.

Two statistical procedures fall under the umbrella of exploratory factor analysis—principal components analysis (PCA) and common factor analysis (ECFA). Given that the purpose of the study was to identify the latent constructs that influence CBM-W, rather than to reduce the variables to a smaller number of linear components, ECFA was deemed most appropriate (Pett et al., 2003; Preacher & MacCallum, 2003). However, preliminary analyses indicated severe multicollinearity. The determinant of the R matrix was less than .00001 ($|R| = 5.25 \text{ E-}15$), and there were a number of extremely high correlations ($r \geq .90$) at each grade level.

To address multicollinearity, trial and error elimination of variables with the highest bivariate correlations was attempted. Elimination of four variables (LegW, WSC, characters, and TWW) produced a determinant greater than .00001 ($|R| = .0000423$), but Haitovsky's test (1969) still indicated that the determinant was not significantly different from zero ($\chi^2(55) = .0105$). Haitovsky's test was only significant when eight variables (LegW, WSC, Characters, TWW, sentences, CIWS, complete sentences, and %CWS) were eliminated from the analysis ($\chi^2(21) = 45.08$).

Although it would have been possible to proceed with ECFA using the remaining seven variables, the analysis would have suffered from significant limitations. Seven variables would only be enough to identify a two factor solution at most, and the results would lack practical and theoretical value

because all of the most commonly used CBM-W indices had been removed from the analysis. Therefore, the decision was made to conduct PCA rather than ECFA. According to Field (2009), multicollinearity does not cause a problem for PCA.

Accordingly, correlation matrices were submitted to principal components analysis (PCA). The solution was iterated two times, because this procedure is less likely to produce Heywood cases (Gorsuch, 2003). One of the critical decisions in factor analysis is the decision regarding how many factors to retain in the model (Fabrigar et al., 1999). The number of factors to retain for rotation was determined by a combination of minimum average partials (MAP) and parallel analysis based on the principal components solution and the 95th percentile criterion (Goldberg & Velicer, 2006; O'Connor, 2000), supplemented by scree test criteria (Fabrigar et al., 1999; Preacher & MacCallum, 2003). Because it is likely that the CBM-W components are interrelated, an oblique rotation method, Promax, was used to search for a simple, parsimonious structure (Fabrigar et al., 1999; Preacher & MacCallum, 2003).

Oblique rotations result in two separate factor matrices—the factor pattern matrix and the factor structure matrix. Both matrices were examined (Henson & Roberts, 2006; Preacher & MacCallum, 2003), but the factor pattern coefficients were the primary focus of interpretation (Gorsuch, 2003). In interpreting the pattern matrix, the guidelines proposed by Stevens (2009) were used. Only coefficients that were both statistically and practically significant were used to interpret a factor. Stevens' recommendation is that an alpha level of .01 (two-

tailed test) should be used and that significance should be determined by doubling the critical value for a normal correlation. Therefore, in the third grade sample coefficients greater than .33 were considered statistically significant, and in the seventh and tenth grade samples coefficients greater than .42 were considered statistically significant. Additionally, Stevens suggested that pattern coefficients greater than or equal to .40 are practically significant. Thus, loadings $\geq .40$ were considered salient for the third grade analysis and loadings $\geq .42$ were considered salient for the seventh and tenth grade analyses.

Chapter 4

Results

Table 7 reports the means and standard deviations for the CBM-W indices. In general, the mean scores appeared to increase with grade level, with the largest differences noted between third and seventh grades. Most variables appeared to have approximately normal distributions, although several of the percentage variables showed a ceiling effect, particularly at the secondary level; and variables such as CPM and sentences showed a floor effect at the third grade level. Ceiling effects were also observed when examining the scatterplots, but there did not appear to be any non-linear relationships. Two cases that produced extreme outliers were identified and excluded from the analysis per Goldberg and Velicer's recommendation (2006), because the scores appeared to be distorting the correlations between %LegW and the other indices. These two participants had written almost their entire responses illegibly.

The correlation matrices for each grade level are presented in Tables 8 through 10. At all three grade levels the majority of variables correlated $\geq |.30|$ with at least three other variables. The exceptions were mean length of T-Units, words per sentence, and %LegW. %LegW had two correlations $\geq |.30|$ at each grade level. Words per sentence had two correlations $\geq |.30|$ in the third grade sample, three correlations of this magnitude in the seventh grade sample, and four in the tenth grade sample. Mean length of T-Units did not have any correlations $\geq |.30|$ in the third grade sample and was considered for elimination from the analyses; however, because mean length of T-Units had two correlations $\geq |.30|$

in the seventh grade sample and three that exceeded that cut-off in the tenth grade sample, the decision was made to retain it.

In addition to examining the correlation matrices, Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin (KMO) test were used to determine whether the correlation matrices were factorable (Pett et al., 2003). Bartlett's Test of Sphericity was significant at all three grade levels, indicating that the correlation matrix was not random (third grade $\chi^2(105) = 7,734.5$; seventh grade $\chi^2(105) = 5,103.5$; tenth grade $\chi^2(105) = 5,051.0$). The Kaiser-Meyer-Olkin (KMO) statistics were .740 for the third grade sample, .699 for the seventh grade sample, and .708 for the tenth grade sample. These values were considered acceptable for factor analysis (Pett et al., 2003).

Parallel analysis, the MAP criteria, and the scree test each indicated three-component solutions for all three samples (third, seventh, and tenth grades). A three-component model accounted for 77.7% of total variance for the third grade sample, 81.0% of the total variance for the seventh grade sample, and 79.6% of total variance for the tenth grade sample. For each analysis three components were rotated using a Promax rotation procedure. Pattern and structure coefficients for the rotated solution are reported in Tables 11 through 13.

The structure of CBM-W indices was fairly consistent across grade levels as quantified by Tucker's congruence coefficient (Tucker, 1951). Congruence coefficients for the first component ranged from .95 to .99 and congruence coefficients for the second component ranged from .97 to .99. Using Lorenzo-Seva and Ten Berge's criterion (2006), the first and second components exhibited

good similarity across all grade levels. The third component also displayed good similarity when the seventh and tenth grade samples were compared ($r_c = .98$) but only fair similarity ($r_c = .92$ and $.93$) when the third grade sample was compared to the other two grades.

At each grade level the first component extracted appeared to be a production component. This finding was consistent with previous factor analyses. At all three grade levels TWW, LegW, WSC, characters, and CLS had high loadings on this component. In the unrotated factor matrix the production component accounted for 44.7% of the total variance for the third grade sample, 47.4% of the variance for the seventh grade sample, and 48.3% of the variance for the tenth grade sample.

All three principal components analyses also identified an accuracy component, similar to other studies. The percentage variables (e.g., %CWS) and ML/CWS loaded highly on this component. The accuracy component was the second component extracted in all three samples. In the unrotated factor matrix this component accounted for 19.3% of the total variance in the third grade sample, 18.1% in the seventh grade sample, and 17.2% in the tenth grade sample.

The third component accounted for 13.7% of the total variance for the third grade sample prior to rotation, 15.5% for the seventh grade sample, and 14.1% for the tenth grade sample. Words per sentence, sentences, complete sentences, CPM, and mean length of T-Units had high loadings on this factor at all three grade levels. Mean length of T-Units is thought to be a measure of sentence complexity and the other indices, words per sentence and complete

sentences, also relate to sentence structure and sentence complexity. Therefore, this component may be described as a complexity component.

Component intercorrelations are reported in Table 14. These correlations ranged from .182 to .381. In the third grade sample all factor correlations were modest and exceeded .30. Tabachnick and Fidell (1983) recommend an oblique rotation in such cases. In the seventh and tenth grade samples intercorrelations were smaller. Therefore, the analyses were run again using an orthogonal rotation, Varimax, to see if it provided a more parsimonious solution. Varimax rotation did not provide simple structure, especially for the seventh and tenth grade samples where it produced more variables with salient loadings on two components (i.e., complex loadings). For that reason it was determined that an oblique rotation was preferred.

Even with oblique rotation, several indices saliently loaded on two components. In all three samples CIWS loaded on both the production and the accuracy components. At the third grade level CIWS loaded primarily on the accuracy component (pattern coefficient = .633), but had a weaker loading on the production component (pattern coefficient = .415). This pattern was reversed at the secondary levels. In the seventh and tenth grade samples CIWS loaded primarily on the production component (seventh grade pattern coefficient = .680, tenth grade pattern coefficient = .739), but had a weaker loading on the accuracy component (seventh grade pattern coefficient = .460, tenth grade pattern coefficient = .491). In the secondary samples sentences loaded primarily on the complexity component (seventh grade pattern coefficient = .760, tenth grade =

.655), but also loaded on the production component (seventh grade = .459; tenth grade = .510). Mean length of T-units had two significant loadings in the third grade sample, and complete sentences and CPM had two significant loadings in the tenth grade sample.

Chapter 5

Discussion

The revision of the Individuals with Disabilities Education Act in 2004 and greater emphasis on the RTI model have led to an increased use of curriculum-based measurement in high stakes decisions. Although the validity of CBM-R is well established, the same cannot be said of CBM-W. To date the research suggests that the validity of CBM-W is moderate at best and further research is needed to establish the validity of these measures (McMaster & Espin, 2007; Parker, Burns, McMaster, & Shapiro, 2012). Establishing the validity of a test is a process that involves examining a variety of validity evidence (Messick, 1995). This study contributes to that process by providing information about the structure of writing as measured by CBM-W indices.

A consistent finding in CBM-W research has been a pattern of decreasing validity coefficients as grade level increases. It was hypothesized that differences in the structure of the writing process as measured by CBM-W may account for this pattern. That is, if CBM-W indices seem to be measuring different aspects of writing at different grade levels, or if the writing process changes as students' writing skills mature, this could lead to construct underrepresentation.

Principal components analyses conducted at grades 3, 7, and 10 indicated a three-component solution at each grade level. The first component was labeled a production component, because the indices that loaded on this component involve to the ability to fluently produce written text (words and letters). TWW, LegW, WSC, characters, and CLS each had high loadings on this component.

This component seems to be related to text production and transcription, which Berninger (1997) has identified as important aspects of the writing process, especially for young and unskilled writers.

The second component was labeled an accuracy component. ML/CWS and the three percentage indices had high loadings on the second component. These indices, especially the percentage indices, have been termed production independent by other authors, because scores on these indices are independent of the length of the writing sample. These indices measure a student's ability to accurately apply conventions, such as correct punctuation, grammar, and spelling.

Thus far the results of the present study align well with the results of previous factor analyses. Although they sometimes applied different labels to the factors, Puranik et al. (2008), Tindal and Parker (1989a, 1991), and Wagner et al. (2011) each found production and accuracy factors in their studies. Therefore, the present study provides confirmation of previous factor analyses.

The third component extracted in the present study was labeled sentence complexity. Sentences, complete sentences, CPM, words per sentence, and mean length of T-units loaded on this component. These indices all seem to relate to sentence construction, or syntax. This is also supported by the fact that mean length of T-units loaded on this factor. Previous factor analytic studies have indicated that mean length of T-units is a measure of syntactic complexity (Puranik et al., 2008; Wagner et al., 2011). Words per sentence and mean length of T-units had negative loadings on this component, whereas the other variables had positive loadings. This means that as the average length of clauses and

sentences increased, the number of sentences and correct punctuation marks decreased. There is a simple explanation for this—if two students write at a similar pace, but one student tends to write longer, more complex sentences, she will produce fewer sentences and fewer ending punctuation marks than her peer in the same time period.

Congruence coefficients indicated that the first two components have good similarity, which suggests that the pattern coefficients can be considered equal across groups. The third component had fair to good similarity across grade levels. These findings do not support the hypothesis that the factor structure of CBM-W indices becomes more complex as students get older, but rather suggest that the structure is largely stable across grade levels. Although the pattern coefficients seemed to be relatively stable across grade levels, component intercorrelations did appear to decrease slightly as grade level increased. This may indicate that the different aspects of writing become more differentiated as writing skills mature.

If the decreasing validity coefficients cannot be explained by differences in the structure of CBM-W indices, this suggests an alternative explanation for the pattern of decreasing criterion-related validity coefficients as grade level increases. If the writing process evolves from primarily a process of transcription and “knowledge-telling” at the elementary level (Berninger et al., 1996, 1997) to a complex process at the secondary level that also involves such things as linguistics (Bradley-Johnson & Lesiak, 1989), planning, organizing, and revising (Hayes, 1996), and yet the CBM-W measures remain static and do not capture

that complexity, they will not have strong validity. Indeed, it seems that the first two CBM-W components--production and accuracy--relate directly to transcription, which Berninger (1997) described as the mechanics of translating thoughts to writing, including not just writing words (production), but also including mechanics such as punctuation and spelling (accuracy).

Although the overall structure did not differ across grade level, there were several differences in the loadings of individual variables. Jewell and Malecki (2005) classified CIWS as an accurate-production index, because they asserted that it is a measure of both writing fluency and accuracy. The results of the present study suggest that their dual classification was accurate. At all three grade levels CIWS had a dual loading on the production and the accuracy components. However, the primary loading did differ by grade level. It appears that at the elementary level CIWS is primarily a measure of accuracy, and secondarily a measure of production. In contrast, at the secondary level CIWS is primarily a measure of production and secondarily a measure of accuracy. This may provide a partial explanation for CIWS's pattern of decreasing criterion-related validity coefficients as grade level increases. If production becomes less important to the writing process at the secondary level and CIWS is primarily measuring production at grades 7 and 10, we would expect the validity coefficients to be lower.

Several of the complexity variables also had dual loadings. In the seventh grade sample sentences had a secondary loading on the production component, and in the tenth grade sample sentences, complete sentences, and CPM all had

secondary loadings on the production component. As was the case with CIWS, this may provide a partial explanation for decreasing validity coefficients at higher grade levels. If these variables are partly measures of production, and production is less critical to the writing process at the secondary level, we would expect validity coefficients to be lower.

Limitations and Future Research

There are several limitations to the present study. Some of these limitations pertain to the study samples. First, all of the participants were recruited from a single southwestern school district. This may limit the generalizability of the findings to other populations. Additionally, the clustered nature of the data (i.e., students were nested within schools) was not considered in the principal components analyses. Another limitation is the relatively small sample sizes. A sample size of 150 participants is near the lower limit of what is considered acceptable for factor analysis.

Kline (1994) recommends replication when sample sizes are small. Therefore, one recommendation for future research is replication of the present study with other samples. This would provide confirmation of the structure indicated by this study, and it would also address the limitation regarding generalizability.

Another limitation of the present study relates to the factor analytic method that was applied. The purpose of the study was to examine the latent constructs underlying CBM-W indices. However, extreme multicollinearity made it necessary to use PCA, rather than ECFA. PCA is a data reduction technique

used to identify a smaller set of composite variables that explain as much of the original variance as possible, but it is not intended to identify latent constructs as is ECFA. The ideal solution would be to conduct an ECFA with CBM-W variables, but this may not be possible even with other samples because it is likely that extreme multicollinearity will be present in those samples as well.

Another direction for future research may be to examine the use of a combination of CBM-W variables for progress monitoring, rather than a single variable. For example, would a combination of variables that measure each of the three components identified in this study—production, accuracy, and complexity—provide better prediction of writing outcomes than any single index? Some research has already been done in this area. Using multiple regression, Amato and Watkins (2011) found that a combination of a complexity variable (CPM) and an accuracy variable (%CWS) provided the best prediction of the TOWL-3 Overall Writing Quotient in an eighth grade sample. The production variables did not explain unique variance in their sample; however, further research is needed in this area to determine which combination of variables may provide the best prediction of outcome variables.

Given existing research and the results of the present study, recommendations can be made regarding combinations of variables that may be most promising. In making these recommendations technical adequacy (reliability and validity), sensitivity to growth across grade level, and ease of administration and scoring were considered.

Among the production variables there is little to distinguish one as preferable to the others. All of the production variables have good reliability (e.g., Espin et al., 2000; Gansle et al., 2006), and in the present study they showed growth across grade levels. In regards to criterion-related validity, the production variables have generally demonstrated moderately strong validity at the elementary level (e.g., Lembke et al., 2003; Tindal & Parker, 1991), whereas at the secondary level the validity coefficients have generally been weaker (e.g., Espin et al., 1999; Jewell & Malecki, 2005; Tindal & Parker, 1989a). Because they are so similar in the other criteria, TWW is recommended because it is the simplest and quickest index to score.

As discussed elsewhere, the accuracy variables are generally not well suited for measuring growth (McMaster & Espin, 2007). In the present study mean scores for the accuracy variables increased from third grade to seventh grade. However, there were no differences in mean scores between seventh and tenth grades. In regards to technical adequacy, there have been mixed findings regarding the reliability of ML/CWS (Espin et al., 2000; Parker et al., 1991b). The percentage measures, on the other hand, have demonstrated acceptable reliability (Parker et al., 1991b; Watkinson & Lee, 1992). Among the percentage measures %CWS has consistently produced the strongest criterion-related validity coefficients (Amato & Watkins; 2011; Jewell & Malecki, 2005; Tindal & Parker, 1989a) and for that reason it is the recommended accuracy variable even though it is more complex to score than %WSC.

In the case of the complexity variables less information is available to guide the recommendation, because these variables have not been studied as extensively as other CBM-W variables. In particular, more evidence is needed to substantiate the reliability of these variables. Among these variables CPM has the strongest validity evidence, and has even produced promising validity coefficients in secondary populations (Diercks-Gransee et al., 2009; Gansle et al., 2004; Gansle et al., 2006). It is also one of the easier complexity variables to score. In regards to sensitivity to growth, the complexity variables seem to suffer from the same weakness as the accuracy variables. In the present study mean scores for the complexity variables increased from third grade to seventh grade, but did not increase from seventh grade to tenth grade. Overall, CPM seems to be the complexity variable with the most potential. Therefore, TWW, %CWS, and CPM are recommended as a promising combination of CBM-W indices for measuring writing across grade levels.

Although there is potential for a combination of variables to provide stronger criterion-related validity than single indices, it may be that strong criterion-related validity can only be achieved if additional indices are discovered that measure other aspects of the writing process, such as organization or vocabulary. This is especially important in the case of secondary students.

Even if a predictive combination of variables is identified, other challenges must be addressed. It is important for CBM-W indices to be time efficient and sensitive to growth (Deno, 2003), but if a combination of variables is needed to achieve adequate validity, it may no longer be time efficient. Also,

using a combination of variables may make it difficult to assess growth. One potential solution to this problem may be the use of computer scoring programs. If programs for scoring writing samples were developed, they could score writing samples efficiently even if several variables were involved. Furthermore, it would not be necessary to limit students' writing to three or five minutes, because a computer program could score a longer writing sample just as quickly as a shorter sample. Use of longer writing samples may give students more time to engage in processes such as planning and organizing, which might provide a more authentic writing assessment. Some research also indicates that longer writing samples produce more reliable and valid results (Espin et al., 2005). One trade-off with this method is that it would be necessary to have students use computers for their writing; therefore, this solution may only be feasible for secondary students.

Conclusion

This study used Principal Components Analysis to examine the structure of writing as measured by CBM-W indices. It was hypothesized that the structure would differ by grade level; however, this was not the case. The overall structure of CBM-W indices was found to remain stable in samples of third, seventh, and tenth grade students. In all cases a three-component solution was supported, with the components being labeled production, accuracy, and sentence complexity. The results support previous factor analyses, which also found production and accuracy components (Puranik et al., 2008; Tindal & Parker, 1989a, 1991; Wagner et al., 2011).

The fact that criterion-related validity coefficients decrease as grade level increases may be explained by the factor structure of CBM-W indices. If the writing process changes as students mature, but CBM-W indices continue to primarily measure transcription this may result in decreased validity. To add to this issue, it appears that several of the CBM-W indices that load on factors other than production in the third grade sample, begin to load on production at the higher grade levels. One potential solution to this problem may be using a combination of indices for progress monitoring, rather than a single index. By selecting a combination of indices that measure each of the three CBM-W components, greater validity may be achieved.

References

- Amato, J. M., & Watkins, M. W. (2011). The predictive validity of CBM writing indices for eighth-grade students. *Journal of Special Education, 44*, 195-204.
- Berninger, V. W. (1999). Coordinating transcription and text generation in working memory during composing: Automatic and constructive processes. *Learning Disability Quarterly, 22*, 99-112.
- Berninger, V. W., Garcia, N. P., & Abbott, R. D. (2008). Multiple processes that matter in writing instruction and assessment. In G. A. Troia (Ed.), *Instruction and assessment for struggling writers: Evidence-based practices* (pp. 15-50). New York, NY: Guilford.
- Berninger, V. W., Vaughan, K. B., Abbott, R. D., Abbott, S. P., Rogan, L. W., Brooks, A., . . . Graham, S. (1997). Treatment of handwriting problems in beginning writers: Transfer from handwriting to composition. *Journal of Educational Psychology, 89*, 652-666.
- Berninger, V., 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, NY: Guilford.
- Child, D. (2006). *The essentials of factor analysis* (3rd ed.). New York, NY: Continuum.
- Christ, T. J., Scullin, S., Tolbize, A., & Jiban, C. L. (2008). Implications of recent research: Curriculum-based measurement of math computation. *Assessment for Effective Intervention, 33*, 198-205.
- CTB/McGraw-Hill. (1985). *California Achievement Test*. Monterey, CA: Author.
- CTB/McGraw-Hill. (1996). *TerraNova*. Monterey, CA: Author.
- CTB MacMillan/McGraw-Hill. (1993). *CTB Writing Assessment*. Monterey, CA: Author.
- Diercks-Gransee, B., Weissenburger, J. W., Johnson, C. L., & Christensen, P. (2009). Curriculum-based measures of writing for high school students. *Remedial and Special Education, 30*, 360-371.
- 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., Mirkin, P., & Marston, D. (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (Vol. IRLD-RR-22). University of Minnesota, Institute for Research on Learning Disabilities.
- 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. A., Scierka, B. J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading & Writing Quarterly*, 15, 5-27.
- Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *Journal of Special Education*, 34, 140-153.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.
- Fewster, S., & MacMillan, P. D. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education*, 23, 149-156.
- Field, A. (2009). *Discovering statistics using SPSS* (3rd Ed.). London, England: Sage.
- Fletcher, J. M., & Vaughn, S. (2009). Response to intervention: Preventing and remediating academic difficulties. *Child Development Perspectives*, 3, 30-37.
- Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measuring in mathematics: A review of the literature. *Journal of Special Education*, 41, 121-139.
- Fuchs, L. S., Deno, S. L., & Marston, D. (1982). *Use of aggregation to improve the reliability of simple direct measures of academic performance* (Vol. IRLD-RR-94). University of Minnesota, Institute for Research on Learning Disabilities.

- Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Effects of instrumental use of curriculum-based measurement to enhance instructional programs. *Remedial and Special Education, 10*, 43-52.
- Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review, 31*, 477-497.
- Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Slider, N. J., Hoffpauir, L. D., Whitmarsh, E. L., & Naquin, G. M. (2004). An examination of the criterion validity and sensitivity to brief intervention of alternate curriculum-based measures of writing skill. *Psychology in the Schools, 41*, 291-300.
- 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.
- Goldberg, L. R., & Velicer, W. F. (2006). Principles of exploratory factor analysis. In S. Strack (Ed.), *Differentiating normal and abnormal personality* (2nd ed., pp. 209-237). New York, NY: Springer.
- Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 143-164). Hoboken, NJ: Wiley.
- Haitovsky, Y. (1969) Multicollinearity in regression analysis: A comment. *Review of Economics and Statistics, 51*, 486-489.
- Hammill, D. D., & Larsen, S. C. (1978). *The Test of Written Language*. Austin, TX: PRO-ED.
- Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test, Ninth Edition*. San Antonio, TX: Author.
- Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1-27). Hillsdale, NJ: Erlbaum.
- Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. Gregg & E. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31-50). Hillsdale, NJ: Erlbaum.

- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66, 393-416.
- 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., . . . Wasileski, T. (1994). Measurement 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, MD: Brookes.
- Hoover, H. D., Hieronymus, A. N., Fisbie, D. A., & Dunbar, S. B. (1996). *Iowa Tests of Basic Skills*. Itasca, IL: Riverside Publishing.
- Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABCs of CBM: A practical guide to curriculum-based measurement*. New York, NY: Guilford.
- Individuals with disabilities education improvement act of 2004. (2004). Retrieved September 18, 2009 from <http://idea.ed.gov/download/statute.html>
- 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.
- Kline, P. (1994). *An easy guide to factor analysis*. London, England: Routledge.
- Lee, L., & Canter, S. M. (1971). Developmental sentence scoring. *Journal of Speech and Hearing Disorders*, 36, 335-340.
- Lembke, E., Deno, S. L., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary school students. *Assessment for Effective Intervention*, 28(3-4), 23-35.
- Lorenzo-Seva, U., & Ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2, 57-64.
- Madden, R., Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1978). *Stanford Achievement Test*. New York, NY: Harcourt Brace Jovanovich.
- Malecki, C. (2008). Best practices in written language assessment and intervention. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 477-488). Bethesda, MD: NASP.

- 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., & Deno, S. (1981). *The reliability of simple, direct measures of written expression*. (Vol. IRLD-RR-50). University of Minnesota, Institute for Research on Learning Disabilities.
- Marston, D., Deno, S. L., & Tindal, G. (1983). *A comparison of standardized achievement tests and direct measurement techniques in measuring pupil progress* (Vol. IRLD-RR-126). University of Minnesota, Institute for Research on Learning Disabilities.
- McCutchen, D. (2006). Cognitive factors in the development of children's writing. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 115-130). New York, NY: Guilford.
- McMaster, K. L., & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review, 37*, 550-556.
- McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing. *Journal of Special Education, 41*, 68-84.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.
- Miller, L. (2009). Informal and qualitative assessment of writing skills in students with disabilities. *Assessment for Effective Intervention, 34*, 178-191.
- Mitzel, H. C., & Borden, C. F. (2000). *LEAP for the 21st century: 1999 operational final technical report*. Monterey, CA: CTB/McGraw-Hill.
- National Center for Education Statistics. (2008). *The nation's report card: Writing 2007*. Retrieved September 18, 2009 from <http://nces.ed.gov/nationsreportcard/pdf/main2007/2008468.pdf>
- National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. Retrieved September 14, 2010 from http://datacenter.spps.org/sites/2259653e-ffb3-45ba-8fd6-04a024ecf7a4/uploads/SOTW_A_Nation_at_Risk_1983.pdf
- National Commission on Writing. (2003). *The neglected "R:" The need for a writing revolution*. Retrieved September 18, 2009 from http://www.writingcommission.org/prod_downloads/writingcom/neglectedr.pdf

- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavioral Research Methods, Instruments, & Computers*, 32, 396-402.
- Office of Special Education Programs. (2006). *28th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act*. Retrieved September 18, 2009 from <http://www.ed.gov/about/reports/annual/osep/2006/parts-b-c/28th-vol-1.pdf>
- Parker, D. C., Burns, M. K., McMaster, K. L., & Shapiro, E. S. (2012). Extending curriculum based assessment to early writing. *Learning Disabilities Research & Practice*, 27, 33-43.
- Parker, R., Tindal, G., & Hasbrouck, J. (1991a). Countable indices of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality*, 2, 1-17.
- Parker, R., Tindal, G., & Hasbrouck, J. (1991b). Progress monitoring with objective measures of writing performance for students with mild disabilities. *Exceptional Children*, 58, 61-73.
- Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. Thousand Oaks, CA: Sage.
- Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, 2, 13-43.
- Puranik, C. S., Lombardino, L. J., & Altmann, L. J. (2008). Assessing the microstructure of written language using a retelling paradigm. *American Journal of Speech-Language Pathology*, 17, 107-120.
- Shaywitz, B. A., Shaywitz, S. E., Blachman, B. A., Pugh, K. R., Fulbright, R. K., Skudlarski, P., . . . Gore, J. C. (2004). Development of left occipitotemporal systems for skilled reading in children after a phonologically-based intervention. *Biological Psychiatry*, 55, 926-933.
- Shinn, M. R., Good, R. H., Knutson, N., & Tilly, W. D. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, 21, 459-479.
- Shinn, M. R., Ysseldyke, J., Deno, S. L., & Tindal, J. (1982). *A comparison of psychometric and functional differences between students labeled learning disabled and low achieving* (Vol. IRLD-RR-71). University of Minnesota, Institute for Research on Learning Disabilities.

- Simos, P. G., Fletcher, J. M., Sarkari, S., Billingsley-Marshall, R., Denton, C. A., & Papanicolaou, A. C. (2007). Intensive instruction affects brain magnetic activity associated with oral word reading in children with persistent reading disabilities. *Journal of Learning Disabilities, 40*, 37-48.
- Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*, 795-819.
- Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York, NY: Routledge.
- Strein, W., Hoagwood, K., & Cohn, A. (2003). School psychology: A public health perspective. *Journal of School Psychology, 41*, 23-38.
- Tabachnick, B. G., & Fidell, L. S. (1983). *Using multivariate statistics*. New York, NY: Harper & Row.
- Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? construct validity of curriculum-based mathematics measures. *School Psychology Review, 31*, 498-513.
- Tindal, G., Germann, G., & Deno, S. L. (1983). *Descriptive research on the Pine County norms: A compilation of findings* (Vol. IRLD-RR-132). University of Minnesota, Institute for Research on Learning Disabilities.
- Tindal, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement* (Vol. IRLD-RR-109). University of Minnesota, Institute for Research on Learning Disabilities.
- Tindal, G., & Parker, R. (1989a). Assessment of written expression for students in compensatory and special education programs. *Journal of Special Education, 23*, 169-183.
- Tindal, G., & Parker, R. (1989b). Development of written retell as a curriculum-based measure in secondary programs. *School Psychology Review, 18*, 328-343.
- Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research & Practice, 6*, 211-218.
- Torgesen, J. K. (2009). The response to intervention instructional model: Some outcomes from a large-scale implementation in reading first schools. *Child Development Perspectives, 3*, 38-40.

- Tucker, L. R. (1951). A method for synthesis of factor analysis studies (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.
- Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice, 18*, 137-146.
- Videen, J., Deno, S. L., & Marston, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression* (Vol. IRLD-RR-84). University of Minnesota, Institute for Research on Learning Disabilities.
- Wagner, R. K., Puranik, C. S., Foorman, B. Foster, E., Wilson, L. G., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. *Reading and Writing, 24*, 203-220.
- 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.
- Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*, 85-120.
- Weissenburger, J. W., & Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *Journal of School Psychology, 43*, 153-169.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised, Tests of Achievement*. Allen, TX: DLM Teaching Resources.

Table 1

Descriptions of Indices in Curriculum-Based Measurement of Writing

Index	Description
Indices introduced by the Institute for Research on Learning Disabilities	
Correct letter sequences (CLS)	Total correct letter sequences. A correct letter sequence is any two adjacent letters that are correct according to the spelling of the word.
Correct writing sequences (CWS)	Also called correct word sequences. A correct writing sequence is any two adjacent words (or a word and a punctuation mark) that are acceptable to a native English speaker within the context of what is written.
Large words (LW)	The total number of words with seven or more letters. Words ending in the suffixes “ed” or “ing” are only counted if the root word was at least seven letters long.
Mature words (MW)	A tally of all words that do not appear on a list of common words—Finn’s Undistinguished Word Choice List.
Mean length of T-units	A T-unit, or minimally terminable unit, is the shortest allowable unit that a sentence can be broken into without becoming a fragment. It can also be described as one main clause with all subordinate clauses attached to it. Mean length of T-units is calculated by counting the total words written and dividing by the number of T-units.
Total words written (TWW)	A count of the number of words in a writing sample. A word is defined as any letter or group of letters separated by a space, regardless of spelling.
Words spelled correctly (WSC)	This index is calculated by subtracting the number of words in the writing sample that are spelled incorrectly from the total words written. A word is incorrectly spelled when it cannot stand alone in the English language.

Index	Description
	Additional indices
Adjectives	The total number of adjectives in a writing sample. Proper adjectives are counted, but demonstrative (e.g., this, that, these) and possessive (e.g., his, hers) adjectives are not counted.
Adverbs	The total number of words modifying a verb within the writing sample.
Characters	Obtained by counting all letters, spaces, and punctuation marks in a writing sample.
Characters per word (C/W)	The number of characters divided by the number of words written.
Complete sentences (CS)	To be counted as a complete sentence a sentence must start with a capital letter, have a recognizable subject, have a verb, and have ending punctuation.
Correct capitalization (CC)	The number of correctly used capitalizations. This includes capitalizations of proper nouns and the first word in each sentence.
Correct minus incorrect writing sequences (CIWS)	Calculated by subtracting the number of incorrect writing sequences from the total number of correct writing sequences.
Correct punctuation marks (CPM)	The total number of correctly applied punctuation marks. To be correct the punctuation mark must be in the correct location in the sentence and be appropriate for the sentence in that location.
Incorrect writing sequences (IWS)	An incorrect writing sequence is counted when one or both words in an adjacent two-word sequence is misspelled, or is syntactically or grammatically unacceptable to a native English speaker.

Index	Description
Legible words (LegW)	The total number of words that are recognizable as English words. Raters view words individually through a mask window, starting at the end of the sample in order to minimize context clues.
Long words	The total number of words spelled correctly in isolation and containing eight or more letters.
Mean length of correct writing sequences (ML/CWS)	Calculated by counting the number of correct writing sequences in a continuous string, summing over all strings, and then dividing by the total number of different strings.
Parts of speech (PS)	The total number of nouns, verbs, and adjectives.
Percentage of correct writing sequences (%CWS)	The percentage of correct writing sequences in a writing sample.
Percentage of legible words (%LW)	The percentage of legible words in a writing sample.
Percentage of words spelled correctly (%WSC)	The percentage of words spelled correctly in a writing sample.
Sentences	Defined as any series of words separated from another series of words by a period, question mark, or exclamation point.
Sentence fragments (SF)	A sentence fragment is an incomplete sentence that cannot stand alone.
Simple sentences (SS)	A simple sentence is an independent clause that contains one subject and one main verb. Only sentences that are complete sentences (as defined above) are counted as simple sentences.
Total punctuation marks (TPM)	The total number of punctuation marks included in the writing sample, regardless of whether they were correctly applied.
Words in complete sentences	The total words in all sentences that meet the criteria for being a complete sentence.

Index	Description
Words per sentence (W/S)	Calculated by counting the total words written and then dividing by the number of sentences.

Table 2

Summary of Reliability Studies Conducted at the Institute for Research on Learning Disabilities

Study	<i>N</i>	Sample	Grade	Type	TWW	WSC	CLS	MW	CWS	T-Units
Marston & Deno (1981, Study 1)	28	LD	1-6	Test-retest	.91, .64	.81, .62	.92, .70	.57, .50		
Marston & Deno (1981, Study 2)	161	GE	1-6	Alternate forms	.95	.95	.96			
76 Marston & Deno (1981, Study 3)	105	GE	1-6	Internal consistency	.87	.70	.87	.74		
Marston & Deno (1981, Study 4)	20	GE	1-6	Interscorer	.98	.98	.99	.92		
Deno et al. (1982)	566	GE	1-6	Interscorer	.96-.99	.96-.99	.96-.99			
Fuchs, Deno, & Marston (1982)	78	LA	3-6	Alternate form		.55-.89				

Study	<i>N</i>	Sample	Grade	Type	TWW	WSC	CLS	MW	CWS	T-Units
Shinn, Ysseldyke, Deno, & Tindal (1982)	71	LD, LA	1-5	Alternate form	.51-.71					
				Test-retest	.69					
Videen, Deno, & Marston (1982)	50	GE	3-6	Interscorer agreement						90.3%
Marston, Deno, & Tindal (1983)	785	LA	3-6	Interscorer	.91-.96	.91-.96	.91-.96			
77 Tindal, Germann, & Deno (1983)	60	GE	4-5	Alternate form	.71		.70			
Tindal, Marston, & Deno (1983)	566	GE	1-6	Alternate form	.73	.72	.93			
				Interscorer	.98	.98	.98			

Note. GE = general education, LD = learning disabilities, LA = low achieving, TWW = total words written, WSC = words spelled correctly, CLS = correct letter sequences, MW = mature words, CWS = correct writing sequences, T-Units = mean length of T-Units.

Table 3

Summary of Validity Studies Conducted at the Institute for Research on Learning Disabilities

Study	N	Sample	Grade	Criterion measure	TWW	WSC	LW	MW	CWS	CLS	T-Units
Deno, Mirkin, & Marston (1980, Study 1)	28	GE, LD	3-6	TOWL	.43-.62	.64-.83		.67			.03-.22 ^a
Deno, Mirkin, & Marston (1980, Study 2)	28	GE, LD	3-6	TOWL	.63-.81	.67-.80	.50-.75	.73-.85			.19 ^a -.60
				SAT	.56-.71	.60-.77	.42 ^b -.72	.52-.77			.03 ^a -.52
Deno, Mirkin, & Marston (1980, Study 3)	82	GE, LD	3-6	DSS	.65-.88	.67-.87	.38-.48	.54-.74		.64-.86	.29
Videen, Deno, & Marston (1982)	50	GE	3-6	DSS					.49		
				TOWL					.69		

Holistic
rating

.85

Note. All coefficients are significant at the .01 level unless otherwise noted. GE = general education, LD = learning disabilities, TWW = total words written, WSC = words spelled correctly, CLS = correct letter sequences, LW = large words, MW = mature words, CWS = correct writing sequences, T-Units = mean length of T-Units, TOWL = *Test of Written Language*, SAT = *Stanford Achievement Test*, DSS = *Developmental Sentence Scoring*.

^a not significant.

^b $p < .05$.

Table 4

Summary of Additional Studies Examining the Technical Adequacy of Curriculum-Based Measurement of Written Expression

Study	N	Sample	Grade	Indices	Criterion validity		Reliability																
					Criterion measure	r	Test-retest	Alternate form	Internal	Inter-scorer													
Elementary Studies																							
Tindal & Parker (1991)	240	GE, LD, LA	3-5	TWW	Analytic rating	-.02-.58					.99												
					SAT							.22											
				WSC	Analytic rating	.13-.63											.97						
					SAT													.28					
				CWS	Analytic rating	.29-.63																	.92
					SAT																		

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability		
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal
Parker, Tindal, & Hasbrouck (1991a, Study 1)	1,917	GE	2-5	TWW	Holistic rating	.36-.49			
				WSC	Holistic rating	.49-.64			
				CWS	Holistic rating	.58-.61			
				%WSC	Holistic rating	.48-.67			
				%CWS	Holistic rating	.43-.70			
Gansle, Noell, VanDerHeyden, Naquin, & Slider (2002)	179	GE	3-4	TWW	ITBS	.15		.62	.96
					LEAP	.16-.28			
					Teacher rankings	.08			

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scorer
				WSC	ITBS	.24		.53		.95
					LEAP	.26-.29				
					Teacher rankings	.21				
				CWS	ITBS	.43		.46		.86
					LEAP	.28-.41				
					Teacher rankings	.36				
				LW	ITBS	.33		.01		.88
					LEAP	.21-.24				
					Teacher rankings	.12				
				CC	ITBS	.26		.43		.92
					LEAP	.15-.18				
					Teacher rankings	.21				

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scorer
				TPM	ITBS	.43		.29		.91
					LEAP	.18				
					Teacher rankings	.32				
				CPM	ITBS	.44		.59		.86
					LEAP	.25-.26				
					Teacher rankings	.37				
				CS	ITBS	.29		.43		.92
					LEAP	.22				
					Teacher rankings	.33				
				SS	ITBS	.38		.44		.71
					LEAP	.01				
					Teacher rankings	.23				

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scoring
				PS	ITBS	.05-.28		.20-.44		.82-.90
					LEAP	.12-.33				
					Teacher rankings	.03-.20				
				W/CS	ITBS	.33		.42		.76
					LEAP	.22-.23				
					Teacher rankings	.33				
				SF	ITBS	.23		-.12		.70
					LEAP	-.12-.11				
					Teacher rankings	.09				
Lembke, Deno, & Hall (2003)	15	LD, LA	2	TWW	Holistic ratings	.06-.62				
					Teacher rankings	.29-.66				

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scoring
				WSC	Holistic ratings		.07-.83			
					Teacher rankings		.13-.73			
				CWS	Holistic ratings		.18-.78			
					Teacher rankings		.25-.73			
				CLS	Holistic ratings		.21-.80			
					Teacher rankings		.38-.65			
				CIWS	Holistic ratings		.56-.84			
					Teacher rankings		.43-.78			
Gansle et al. (2004)	45	GE	3-4	TWW	WJ-R		.23			.99

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scorer
Gansle, VanDerHeyden, Noell, Resetar, & Williams (2006)	538	GE	1-5	CWS	WJ-R	.36				.91
				CPM	WJ-R	.34				.97
				TPM	WJ-R	.42				.99
				SS	WJ-R	-.05				.78
				W/CS	WJ-R	.35				.67
				TWW	SAT	.34	.80			.98
				WSC	SAT	.38	.82			.97
				CWS	SAT	.43	.78			.94
				CPM	SAT	.39	.64			.91
				CC	SAT	.28	.44			.94
CS	SAT	.36	.65			.84				

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scorer
				W/CS	SAT	.41	.61			.82
Junior High Studies										
Tindal & Parker (1989a)	172	LD, LA	6-8	TWW	Holistic ratings	.10				.99
				WSC	Holistic ratings	.31				.98
				CWS	Holistic ratings	.45				.87
				LegW	Holistic ratings	.24				.95
				ML/CWS	Holistic ratings	.59				.83
				%WSC	Holistic ratings	.73				.98
				%CWS	Holistic ratings	.75				.87

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scoring
Parker, Tindal, & Hasbrouck (1991b)	36	LD	6-8	%LegW	Holistic ratings	.42				.92
				TWW	TOWL	.16	.69-.83		.77	.99
					Holistic ratings	.39				
				WSC	TOWL	.25	.68-.79		.78	.97
					Holistic ratings	.54				
				CWS	TOWL	.27	.49-.77		.75	.87
					Holistic ratings	.64				
				LegW	TOWL	.26	.69-.83		.81	.95
					Holistic ratings	.45				

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scoring
				ML/CWS	TOWL	.18	.26-.66		.78	.97
					Holistic ratings	.63				
				%WSC	TOWL	.28	.45-.75		.77	.89
					Holistic ratings	.53				
				%LegW	TOWL	.56	.17-.76		.79	.92
					Holistic ratings	.60				
Watkinson & Lee (1992)	52	GE, LD	6-8	TWW					.99	
				WSC					.96	
				CWS					.95	
				LegW					.97	
				IWS					.87	

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability		
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal
				%WSC					.80
				%CWS					.82
				%LegW					.87
Espin et al. (2000)	112	GE, LD	7-8	TWW	District test	.43-.47		.73-.77	1.00
					Holistic ratings	.34-.46			
				WSC	District test	.46-.51		.72-.76	1.00
					Holistic ratings	.38-.48			
				CWS	District test	.61-.65		.75-.80	
					Holistic ratings	.54-.60			

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scoring
				ML/CWS	District test			.32-.57		.86
					Holistic ratings					
				CIWS	District test	.69-.75		.72-.78		.88
					Holistic ratings	.65-.70				
				Characters	District test	.47-.51		.78-.81		1.00
					Holistic ratings	.40-.50				
				C/W	District test			.12-.47		1.00
					Holistic ratings					

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scoring
				Sentences	District test	.66-.77		.61-.82		1.00
					Holistic ratings	.54-.64				
				W/S	District test	-.76 to -.61		.61-.80		1.00
					Holistic ratings	-.39-.37				
Fewster & MacMillan (2002)	465	GE	6-7	WSC	Eng 8	.34				
					Eng 9	.29				
					Eng 10	.28				
Espin, De La Paz, Scierka, & Roelofs (2005)	22	GE, LD	7-8	TWW	Holistic ratings	.58-.82				
					Functional elements	.68-.90				

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability		
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal
Amato & Watkins (2011)	447	GE	8	CWS	Holistic ratings	.68-.83			
					Functional elements	.70-.79			
				CIWS	Holistic ratings	.67-.82			
					Functional elements	.66-.70			
				TWW	TOWL	.34			
				WSC	TOWL	.37			
				CWS	TOWL	.49			
CIWS	TOWL	.56							
	Sentences	TOWL	.28						
	CC	TOWL	.23						

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability		
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal
				TPM	TOWL	.43			
				CPM	TOWL	.44			
				%WSC	TOWL	.41			
				%CWS	TOWL	.61			
High School Studies									
94 Espin, Scierka, Skare, & Halverson (1999)	147	GE, LD, LA	10	TWW	CAT	.13			
					Eng GPA	.22-.25			
					Holistic ratings	.36			
				WSC	CAT	.17			
					Eng GPA	.25-.29			
					Holistic ratings	.41			

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scorer
				CWS	CAT	.29				
					Eng GPA	.33-.35				
					Holistic ratings	.52				
				ML/CWS	CAT	.34				
					Eng GPA	.20-.23				
					Holistic ratings	.40				
				Characters	CAT	.24				
					Eng GPA	.33-.36				
					Holistic ratings	.48				
				C/W	CAT	.41				
					Eng GPA	.32-.36				
					Holistic ratings	.38				

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scoring
96 Diercks-Gransee, Weissenburger, Johnson, Christensen (2009)	82	GE, LD	10	Sentences	CAT	.40				
					Eng GPA	.43-.45				
					Holistic ratings	.63				
				CPM	WKCE	.28		.76		
					Holistic ratings	.62				
					IWS	WKCE	-.51		.75	
				Adjectives	Holistic ratings	-.71				
					WKCE	.19		.17		
					Holistic ratings	.18				
				Adverbs	WKCE	.01		.14		
					Holistic ratings	.21				

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability		
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal
Studies Across Grade Levels									
Tindal & Parker (1989b)	267	GE	6,8,11	TWW	Holistic rating	-.13-.28			
Parker, Tindal, & Hasbrouck (1991a, Study 2)	243	GE	6,8,11	TWW	Holistic rating	.39-.43			
				WSC	Holistic rating	.43-.52			
				CWS	Holistic rating	.48-.56			
				%WSC	Holistic rating	.34-.46			
				%CWS	Holistic rating	.36-.42			

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability		
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal
Jewell & Malecki (2005)	203	GE	2,4,6	TWW	SAT	-.14-.24			
					THASS	.16-.44			
					LA	.12-.45			
				WSC	SAT	-.05-.38			
					THASS	.24-.49			
					LA	.20-.51			
				CWS	SAT	.23-.57			
					THASS	.46-.58			
					LA	.30-.59			
				CIWS	SAT	.41-.62			
					THASS	.54-.56			
					LA	.36-.61			
				%WSC	SAT	.46-.50			
					THASS	.34-.39			
					LA	.45-.53			

Study	<i>N</i>	Sample	Grade	Indices	Criterion validity		Reliability				
					Criterion measure	<i>r</i>	Test-retest	Alternate form	Internal	Inter-scorer	
Weissenburger & Espin (2005)	484	GE	4,8,10	%CWS	SAT	.52-.67					
					THASS	.40-.49					
					LA	.29-.58					
				TWW	WKCE	.04-.54		.55-.84			
					CWS	WKCE	.18-.62		.59-.84		
					CIWS	WKCE	.29-.68		.61-.82		
McMaster & Campbell (2008)	122	GE	3,5,7	TWW	TOWL	ns		.51-.91			
					MCA	ns					
					LA	ns					
					WSC	TOWL	ns-.60		.52-.90		
						MCA	ns-.45				
						LA	ns-.53				

Study	N	Sample	Grade	Indices	Criterion validity		Reliability			
					Criterion measure	r	Test-retest	Alternate form	Internal	Inter-scoring
				CWS	TOWL	ns-.70		.54-.93		
					MCA	ns-.56				
					LA	ns-.62				
				CIWS	TOWL	ns-.70		.55-.91		
					MCA	.54-.68				
					LA	ns-.72				

Note. GE = general education; LD = learning disabilities; TWW = total words written; WSC = words spelled correctly; CLS = correct letter sequences; LW = large words; MW = mature words; CWS = correct writing sequences; ML/CWS = mean length of correct writing sequence; IWS = incorrect writing sequences; T-Units = mean length of T-Units; CC = correct capitalization; TPM = total punctuation marks; CPM = correct punctuation marks; CS = complete sentences; SS = simple sentences; PS = parts of speech; W/CS = words in complete sentences; SF = sentence fragments; CIWS = correct writing sequences minus incorrect writing sequences; LegW = legible words; %WSC = percent of words spelled correctly; %CWS = percent of correct writing sequences; %LegW = percent of legible words; C/W = characters per word; W/S = words per sentence; GPA = grade point average, EN8, EN9, EN10 = English GPA for grades 8, 9, and 10; LA = language arts GPA; CAT = *California Achievement Test*; TOWL = *Test of Written Language*; SAT = *Stanford Achievement Test*; DSS = *Developmental Scoring System*; MCA = *Minnesota Comprehensive Assessment*; THASS = *Tindal & Hasbrouck Analytic Scoring System*; WKCE = *Wisconsin Knowledge and Concepts Exam*; WJ-R = *Woodcock-Johnson-Revised* writing samples subtest.

ns = not significant.

Table 5

Sample Demographics by Grade Level

	3rd Grade		7th Grade		10th Grade	
	<i>(n = 253)</i>		<i>(n = 154)</i>		<i>(n = 154)</i>	
	<i>n</i>	<i>%</i>	<i>n</i>	<i>%</i>	<i>n</i>	<i>%</i>
Gender						
Male	112	44.3	70	45.5	63	40.9
Female	141	55.7	84	54.5	91	59.1
Ethnicity						
Caucasian	163	64.4	85	55.2	98	63.6
Hispanic	48	19.0	40	26.0	32	20.8
Asian/Pacific Islander	32	12.6	21	13.6	12	7.8
African American	7	2.8	8	5.2	11	7.1
Native American	1	.4			1	.6
Multiracial	2	.8				
Special Education						
Yes	33	13.0	6	3.9	12	7.8
No	220	87.0	148	96.1	142	92.2
Age (in months)						
<i>M</i>	110.8		155.7		192.8	
<i>SD</i>	4.36		5.93		4.85	

Table 6

*Interscorer Reliability for CBM-W Indices: Percent Agreement Between Scorers
and Primary Investigator*

	Percent agreement
Legible Words (LegW)	99.0
Characters	99.3
Total Words Written (TWW)	99.6
Words Spelled Correctly (WSC)	99.3
Sentences	96.7
Complete Sentences	90.2
Correct Punctuation Marks (CPM)	95.0
Correct Minus Incorrect Writing Sequences (CIWS)	94.3
Correct Writing Sequences (CWS)	97.5
Incorrect Writing Sequences (IWS)	80.1
Strings of correct writing sequences	84.2
Correct Letter Sequences (CLS)	98.7
T-Units	98.5

Table 7

Means and Standard Deviations for CBM-W Indices at Three Grade Levels

	3rd Grade		7th Grade		10th Grade	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Characters	183.8	60.7	314.5	76.7	366.2	83.3
Complete sentences (CS)	2.6	2.0	4.8	1.9	4.3	1.9
Correct minus incorrect writing sequences (CIWS)	24.9	13.9	55.2	17.9	61.1	19.0
Correct punctuation marks (CPM)	4.5	3.1	8.0	3.6	8.3	3.9
Correct letter sequences (CLS)	176.2	58.7	305.8	74.6	355.4	80.5
Legible words (LegW)	37.7	12.5	61.1	14.8	68.9	15.6
Mean length of correct writing sequences (ML/CWS)	10.0	7.6	25.5	19.6	25.6	21.9
Mean length of T-unit	7.3	1.9	9.2	2.2	10.9	2.9
Percentage of correct writing sequences (%CWS)	79.4	13.1	91.1	8.5	90.7	8.6
Percentage of legible words (%LW)	99.1	2.1	99.4	1.2	98.7	2.3
Percentage of words spelled correctly (%WSC)	95.2	4.6	97.7	3.0	97.9	2.9
Sentences	3.2	2.0	5.2	1.7	5.0	1.8
Total words written (TWW)	38.0	12.6	61.5	14.9	69.8	15.8
Words per sentence (W/S)	13.0	10.5	12.7	4.1	15.0	5.0
Words spelled correctly (WSC)	36.2	12.2	60.1	14.8	68.3	15.8

Table 8

Third Grade Sample: Correlation Matrix for Curriculum Based Writing Indices (n = 253)

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Sentences	-													
2. CS	.931**	-												
3. CIWS	.601**	.609**	-											
4. CPM	.819**	.802**	.548**	-										
5. T-Units	-.190*	-.127	.153	-.219**	-									
6. W/S	-.325**	-.393**	-.061	-.289**	.040	-								
7. ML/CWS	.288**	.375**	.646**	.316**	.164*	-.109	-							
8. %WSC	.124	.172*	.556**	.193*	.170*	.001	.488**	-						
9. %CWS	.348**	.409**	.792**	.373**	.176*	-.089	.751**	.684**	-					
10. %LegW	.087	.094	.285**	.079	.037	.020	.224**	.518**	.330**	-				
11. CLS	.463**	.411**	.624**	.388**	.091	.075	.135	.129	.102	.090	-			
12. Characters	.487**	.430**	.620**	.420**	.078	.065	.120	.104	.091	.070	.993**	-		
13. TWW	.451**	.390**	.609**	.361**	.091	.086	.098	.107	.074	.071	.971**	.978**	-	
14. LegW	.455**	.395**	.624**	.366**	.093	.086	.111	.136	.093	.122	.972**	.977**	.998**	-
15. WSC	.463**	.410**	.674**	.385**	.109	.085	.162*	.233**	.162*	.136	.967**	.970**	.991**	.994**

Note. ** $p \leq .001$, * $p \leq .01$; TWW = total words written; WSC = words spelled correctly; CLS = correct letter sequences; ML/CWS = mean length of correct writing sequence; T-Units = mean length of T-Units; CPM = correct punctuation marks; CS = complete sentences; CIWS = correct writing

sequences minus incorrect writing sequences; LegW = legible words; %WSC = percent of words spelled correctly; %CWS = percent of correct writing sequences; %LegW = percent of legible words; W/S = words per sentence.

Table 9

Seventh Grade Sample: Correlation Matrix for Curriculum Based Writing Indices (n = 154)

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Sentences	-													
2. CS	.932**	-												
3. CIWS	.603**	.635**	-											
4. CPM	.656**	.635**	.584**	-										
5. T-Units	-.306**	-.246*	.135	-.216*	-									
6. W/S	-.592**	-.601**	.003	-.291**	.492**	-								
7. ML/CWS	.089	.205	.553**	.254**	.117	.076	-							
8. %WSC	.145	.246*	.556**	.255**	.092	.064	.523**	-						
9. %CWS	.251*	.405**	.674**	.348**	.026	-.116	.698**	.791**	-					
10. %LegW	.122	.196	.197	.155	-.036	.020	.239*	.394**	.352**	-				
11. CLS	.541**	.458**	.779**	.490**	.216*	.176	.209*	.187	.185	.028	-			
12. Characters	.560**	.477**	.772**	.513**	.196	.170	.192	.144	.154	.028	.990**	-		
13. TWW	.565**	.485**	.797**	.468**	.218*	.215*	.193	.139	.138	.053	.934**	.949**	-	
14. LegW	.566**	.491**	.803**	.474**	.217*	.214*	.204	.155	.151	.096	.932**	.946**	.999**	-
15. WSC	.567**	.504**	.848**	.489**	.226*	.212*	.253*	.248*	.224*	.089	.934**	.944**	.993**	.994**

Note. ** $p \leq .001$, * $p \leq .01$; TWW = total words written; WSC = words spelled correctly; CLS = correct letter sequences; ML/CWS = mean length of correct writing sequence; T-Units = mean length of T-Units; CPM = correct punctuation marks; CS = complete sentences; CIWS = correct writing

sequences minus incorrect writing sequences; LegW = legible words; %WSC = percent of words spelled correctly; %CWS = percent of correct writing sequences; %LegW = percent of legible words; W/S = words per sentence

Table 10

Tenth Grade Sample: Correlation Matrix for Curriculum Based Writing Indices (n = 152)

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Sentences	-													
2. CS	.891**	-												
3. CIWS	.621**	.642**	-											
4. CPM	.717**	.621**	.590**	-										
5. T-Units	-.302**	-.216*	-.041	-.307**	-									
6. W/S	-.591**	-.556**	-.073	-.365**	.426**	-								
7. ML/CWS	.201	.361**	.610**	.259**	-.056	-.042	-							
8. %WSC	.035	.066	.393**	.152	-.026	-.022	.365**	-						
9. %CWS	.117	.278**	.631**	.209*	-.057	-.021	.607**	.724**	-					
10. %LegW	-.176	-.119	.116	-.091	-.017	.066	.170	.389**	.388**	-				
11. CLS	.625**	.549**	.771**	.551**	.125	.023	.269**	-.012	.078	-.157	-			
12. Characters	.641**	.559**	.757**	.575**	.107	.003	.252*	-.049	.050	-.183	.998**	-		
13. TWW	.668**	.567**	.776**	.552**	.050	-.017	.241*	-.034	.039	-.178	.964**	.965**	-	
14. LegW	.656**	.558**	.797**	.548**	.049	-.010	.263**	.006	.077	-.069	.960**	.957**	.994**	-
15. WSC	.669**	.570**	.819**	.569**	.045	-.019	.290**	.078	.117	-.127	.959**	.955**	.993**	.992**

Note. ** $p \leq .001$, * $p \leq .01$; TWW = total words written; WSC = words spelled correctly; CLS = correct letter sequences; ML/CWS = mean length of correct writing sequence; T-Units = mean length of T-Units; CPM = correct punctuation marks; CS = complete sentences; CIWS = correct writing

sequences minus incorrect writing sequences; LegW = legible words; %WSC = percent of words spelled correctly; %CWS = percent of correct writing sequences; %LegW = percent of legible words; W/S = words per sentence.

Table 11

*Component Loadings for CBM-W Indices for Sample of Third Grade Students:**Principal Components Analysis with Promax Rotation*

CBM-W Index	Pattern (Structure) Coefficients			Communality
	Production	Accuracy	Complexity	
TWW	1.030* (.989)	-.081 (.225)	-.040 (.321)	.987
LegW	1.024* (.991)	-.047 (.255)	-.048 (.324)	.987
Characters	1.007* (.988)	-.084 (.237)	.017 (.369)	.981
CLS	1.007* (.984)	-.051 (.256)	-.020 (.344)	.971
WSC	1.001* (.991)	.038 (.329)	-.057 (.338)	.984
W/S	.333 (.100)	.092 (-.064)	-.688* (-.526)	.390
Sentences	.195 (.516)	-.023 (.363)	.861* (.927)	.890
CS	.108 (.459)	.065 (.427)	.869* (.935)	.890
CPM	.106 (.433)	.034 (.381)	.831* (.884)	.793
T-Units	.179 (.110)	<u>.426*</u> (.281)	-.533* (-.304)	.301
CIWS	<u>.415*</u> (.681)	.633* (.830)	.178 (.575)	.911
%WSC	-.038 (.172)	.907* (.823)	-.191 (.136)	.714
%CWS	-.179 (.159)	.924* (.917)	.129 (.410)	.872
ML/CWS	-.136 (.161)	.792* (.799)	.130 (.378)	.660
%LegW	-.002 (.115)	.615* (.540)	-.197 (.034)	.325

*Statistically and practically significant using Stevens' (2009) criterion of $\geq .40$.

Note. Bold values indicate the primary salient loading. Underlined values indicate a variable that had a salient loading on a second component. TWW = total words written; WSC = words spelled correctly; CLS = correct letter sequences; ML/CWS = mean length of correct writing sequence; T-Units = mean length of T-Units; CPM = correct punctuation marks; CS = complete sentences; CIWS = correct writing sequences minus incorrect writing sequences; LegW = legible words; %WSC = percent of words spelled correctly; %CWS = percent of correct writing sequences; %LegW = percent of legible words; W/S = words per sentence.

Table 12

*Component Loadings for CBM-W Indices for Sample of Seventh Grade Students:**Principal Components Analysis with Promax Rotation*

CBM-W Index	Pattern (Structure) Coefficients			Communality
	Production	Accuracy	Complexity	
TWW	1.047* (.979)	-.114 (.258)	-.089 (.172)	.980
LegW	1.037* (.979)	-.089 (.279)	-.088 (.176)	.975
Characters	1.028* (.968)	-.116 (.257)	-.058 (.198)	.954
CLS	1.011* (.960)	-.076 (.284)	-.084 (.177)	.934
WSC	1.015* (.985)	-.007 (.351)	-.097 (.181)	.980
W/S	.358 (.121)	.071 (-.026)	-.955* (-.838)	.842
Sentences	<u>.459*</u> (.643)	-.069 (.288)	.760* (.870)	.936
CS	.335 (.585)	.121 (.427)	.740* (.861)	.885
CPM	.383 (.579)	.151 (.417)	.502* (.644)	.608
T-Units	.359 (.200)	.125 (.079)	-.743* (-.614)	.538
CIWS	.680* (.867)	<u>.460*</u> (.729)	.053 (.353)	.944
%WSC	-.076 (.241)	.919* (.865)	-.101 (.101)	.767
%CWS	-.115 (.265)	.962* (.935)	.069 (.271)	.888
ML/CWS	.004 (.271)	.824* (.788)	-.154 (.047)	.643
%LegW	-.135 (.084)	.564* (.520)	.027 (.126)	.285

*Statistically and practically significant using Stevens' (2009) criterion of $\geq .40$.

Note. Bold values indicate the primary salient loading. Underlined values indicate a variable that had a salient loading on a second component. TWW = total words written; WSC = words spelled correctly; CLS = correct letter sequences; ML/CWS = mean length of correct writing sequence; T-Units = mean length of T-Units; CPM = correct punctuation marks; CS = complete sentences; CIWS = correct writing sequences minus incorrect writing sequences; LegW = legible words; %WSC = percent of words spelled correctly; %CWS = percent of correct writing sequences; %LegW = percent of legible words; W/S = words per sentence.

Table 13

*Component Loadings for CBM-W Indices for Sample of Tenth Grade Students:**Principal Components Analysis with Promax Rotation*

CBM-W Index	Pattern (Structure) Coefficients			Communality
	Production	Accuracy	Complexity	
TWW	1.037* (.977)	-.113 (.096)	-.090 (.287)	.976
LegW	1.025* (.973)	-.047 (.156)	-.110 (.275)	.960
Characters	1.042* (.972)	-.111 (.093)	-.120 (.260)	.971
CLS	1.045* (.971)	-.072 (.127)	-.152 (.236)	.968
WSC	1.019* (.980)	-.014 (.191)	-.093 (.295)	.970
W/S	.262 (-.085)	.092 (-.025)	-.955* (-.838)	.776
Sentences	<u>.510*</u> (.741)	-.091 (.139)	.655* (.834)	.911
CS	<u>.422*</u> (.675)	.066 (.271)	.620* (.794)	.796
CPM	<u>.442*</u> (.650)	.066 (.254)	.505* (.686)	.650
T-Units	.337 (.040)	-.007 (-.073)	-.769* (-.641)	.507
CIWS	.739* (.851)	<u>.491*</u> (.655)	.015 (.388)	.956
%WSC	-.128 (.049)	.850* (.818)	-.021 (.085)	.687
%CWS	-.045 (.169)	.935* (.930)	.026 (.179)	.867
ML/CWS	.198 (.350)	.664* (.710)	.020 (.217)	.545
%LegW	-.234 (-.152)	.637* (.559)	-.149 (-.122)	.409

*Statistically and practically significant using Stevens' (2009) criterion of $\geq .40$.

Note. Bold values indicate the primary salient loading. Underlined values indicate a variable that had a salient loading on a second component. TWW = total words written; WSC = words spelled correctly; CLS = correct letter sequences; ML/CWS = mean length of correct writing sequence; T-Units = mean length of T-Units; CPM = correct punctuation marks; CS = complete sentences; CIWS = correct writing sequences minus incorrect writing sequences; LegW = legible words; %WSC = percent of words spelled correctly; %CWS = percent of correct writing sequences; %LegW = percent of legible words; W/S = words per sentence.

Table 14


*Component Intercorrelations for Principal Components Analysis with Promax**Rotation*

	Accuracy	Complexity
Third grade		
Production	.312	.381
Accuracy		.377
Seventh grade		
Production	.376	.276
Accuracy		.243
Tenth grade		
Production	.218	.384
Accuracy		.182

APPENDIX A

INSTITUTIONAL REVIEW BOARD/HUMAN SUBJECTS APPROVAL

To: Marley Watkins
EDUC - I.

From: for Mark Roosa, Chair 
Soc Beh IRB

Date: 04/08/2011

Committee Action: Expedited Approval

Approval Date: 04/08/2011

Review Type: Expedited F7

IRB Protocol #: 1103006230

Study Title: The Factor Structure of Curriculum-Based Writing Indices at Grades 3, 7, 10

Expiration Date: 04/03/2012

The above-referenced protocol was approved following expedited review by the Institutional Review Board.

It is the Principal Investigator's responsibility to obtain review and continued approval before the expiration date. You may not continue any research activity beyond the expiration date without approval by the Institutional Review Board.

Adverse Reactions: If any untoward incidents or severe reactions should develop as a result of this study, you are required to notify the Soc Beh IRB immediately. If necessary a member of the IRB will be assigned to look into the matter. If the problem is serious, approval may be withdrawn pending IRB review.

Amendments: If you wish to change any aspect of this study, such as the procedures, the consent forms, or the investigators, please communicate your requested changes to the Soc Beh IRB. The new procedure is not to be initiated until the IRB approval has been given.

Please retain a copy of this letter with your approved protocol.

APPENDIX B

CBM-W ADMINISTRATION INTEGRITY SELF-CHECK

Accuracy of Implementation Rating Scale
(Adapted from AIMSWEB AIRS-WE-CBM)
CBM – Written Expression (CBM-W)

X = Completed accurately, 0 = Completed inaccurately

Testing Procedure	Observation			
	1	2	3	4
Provides students with a pencil and lined sheet of paper. <i>Give each student a copy of the response sheet, and insure that each student has a pencil to write with. Have them write their names in the designated space.</i>	—	—	—	—
Says standardized instructions verbatim. <i>“You are going to write a story. First I will read a sentence, and then you will write a story about what happens next. You will have 1 minute to think about what you will write, and 3 minutes to write your story. Remember to do your best work. If you don’t know how to spell a word, you should guess. Are there any questions?”</i>	—	—	—	—
Says, <i>“Put your pencils down and listen. For the next minute think about ‘One day our teacher was sick. We had a substitute teacher and . . .’”</i>	—	—	—	—
Starts stopwatch.	—	—	—	—
Provides prompt at 30 seconds into one minute think time. <i>“You should be thinking about ‘One day our teacher was sick. We had a substitute teacher and...’”</i>	—	—	—	—
Stops stopwatch at the end of one minute.	—	—	—	—
Says, “Now begin writing.” and restarts stopwatch	—	—	—	—
Provides prompt at 90 seconds into 3 minute writing time <i>“You should be writing about ‘One day our teacher was sick. We had a substitute teacher and . . .’”</i>	—	—	—	—
Monitors student attention to task—gives encouragement/prompts if student stops writing or is looking around.	—	—	—	—
Times for 3 minutes	—	—	—	—
Says, “Stop. Put down your pencil.”	—	—	—	—
Stops stopwatch.	—	—	—	—
Collects all papers and checks to make sure that all students wrote their name.	—	—	—	—

--	--	--	--

Additional Comments: