2014

Multivariate Generalizability of Writing Curriculum-Based Measurement (CBM): An Examination of Form, Occasion, and Scoring Method

Katherine Hunter Chenier
Louisiana State University and Agricultural and Mechanical College, katherinehchenier@gmail.com

MULTIVARIATE GENERALIZABILITY OF WRITING CURRICULUM-BASED MEASUREMENT (CBM): AN EXAMINATION OF FORM, OCCASION, AND SCORING METHOD

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Psychology

by
Katherine H. Chenier
B.A., Tufts University, 2008
M.A., Louisiana State University, 2012
August 2014

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Curriculum-based measurement (CBM) is an assessment technique that has become increasingly popular in schools, gaining importance with the recent national emphasis on school and teacher accountability for student achievement. CBM is used to monitor student performance to provide an indicator of which students are at-risk of not achieving grade level standards and thus are in need of intervention. CBM is easy to administer, utilizes standard procedures, and provides measures indicative of general achievement in various domains. The utility of CBM to measure student ability in writing has been well-established. However, there is a paucity of technical adequacy research for writing CBM compared to CBM in reading and math. Additionally, various scoring methods for writing CBM have been proposed and tested with variable results. This study investigated the reliability of writing CBM using multivariate generalizability (G) theory. The dependability of the measure across forms and occasions for a composite dependent variable consisting of 7 different scoring methods was investigated. Additionally, univariate G theory studies were conducted for each individual scoring method. Results suggested that a composite measure and all independent measures are dependable when 3 forms are administered on 3 occasions for students in grades 3-5, with person contributing the most variance. Additionally, support was found for the use of a composite measure, TWW, WSC, CWS, and CIWS for screening and progress monitoring purposes with 2 forms administered on 1 occasion.

Keywords: curriculum-based measurement, CBM, writing, generalizability theory

INTRODUCTION

The No Child Left Behind Act (NCLB) of 2001 called for increased educational accountability through the measurement of annual yearly progress (AYP) scores connected to improvement goals for schools and districts. According to the act, schools that fail to meet improvement goals can be subject to punishment and may be taken over by state districts (Gresham, Reschly, & Shinn, 2010). For many states, including Louisiana, AYP scores are predominantly based on student performance on end of the year standardized tests. These tests typically cover a broad range of material, with an emphasis on reading, math, and writing (Louisiana Department of Education, 2011). Part of the increased accountability for schools involves systematic and ongoing measurement of student growth in these core subject areas. Academic achievement is required to be tested repeatedly throughout the year with results documented as an indication of student progress. These results are then used to make instructional decisions and provide intervention to those who need it. This repeated documentation is typically done through a method called progress monitoring (Gansle, VanDerHeyden, Noell, Resetar, & Williams 2006; Shapiro, Hilt-Panahon, & Gischlar, 2010). According the website of the National Center on Student Progress Monitoring (2012), progress monitoring is defined as "a scientifically based practice that is used to assess students' academic performance and evaluate the effectiveness of instruction," which involves continuous, repeated assessment of student progress ("What is Progress Monitoring?" para. 1). Progress monitoring has also gained relative importance in schools with the 2004 re-authorization of the Individuals with Disabilities Education Improvement Act (IDEIA), which allowed for the use of a Response to Intervention (RTI) model for identifying students at-risk for learning disabilities. Part of an

RTI model involves screening and progress monitoring of student performance to allow for efficient identification of students in need of more intensive services (Deno et al., 2009).

Reading, math, and writing skills have importance beyond school performance scores and standardized tests. The focus of the current study is writing, which is a core part of the school curriculum and a primary mechanism through which students are expected to express their ideas and knowledge. The ability to write well becomes increasingly important as students advance into college-level education and begin careers (National Commission of Writing, 2003). According to one recent survey, over 90% of professionals deemed effective writing as essential to their work (Light, 2001 as cited in National Commission of Writing, 2003). Additionally, the importance of writing extends beyond school and career. In today's society, electronic writing, such as email, texting, and blog forums, is a daily part of most people's lives and dominates interpersonal communication (National Commission on Writing, 2008; Olinghouse & Santangelo, 2010). A specific writing portion was added to the latest version of the SAT® in 2005, making writing skills even more critical for student's seeking higher education (National Commission on Writing, 2003).

According to the latest statistics from the National Center for Education Statistics (NCES; 2008) of the United States Department of Education, in 2007 only 24% of twelfth grade students and 33% of eight grade students were proficient in writing and in 2002 only 28% of fourth grade students were proficient. These statistics suggest that more time devoted to writing is needed in schools. The National Commission on Writing (2003) has deemed writing the "most neglected" of the three "Rs" (writing, reading, and arithmetic; pg 3). The commission also has highlighted the importance of assessing student writing in a manner that involves actual writing samples, versus simply relying on multiple-choice items or items that can be scored by a

machine. Given the vast importance of writing skills, it is critical to have a technically adequate measure for assessing student writing, as well as a measure for progress monitoring. A progress monitoring technique that has become increasingly popular over the last few decades and has been suggested for use within a standards-based accountability system is an assessment method known as curriculum-based measurement (CBM; Quenemoen, Thurlow, Moen, Thompson, & Morse, 2004). Curriculum-based measurement (CBM) has its roots in Data-Based Program Modification (DBPM).

Data-Based Program Modification

Curriculum-based measurement (CBM) grew out of a program of study conducted by Stanley Deno and Phyllis Mirkin at the University of Minnesota in the 1970s called Data-Based Program Modification (DBPM; Hosp, Hosp, & Howell, 2007; Lai, Park, Anderson, Alonzo, & Tindal, 2012; Shinn, 2010). Deno and Mirkin created DBPM as a way to have a standardized system of data collection that could be administered repeatedly in order to guide intervention and instructional modification, specifically in regards to Individualized Education Plans (IEPs) for students in special education. With the passage of the Education for All Handicapped Children Act (later becoming the Individuals with Disabilities Education Act; PL 94-142, 1975), all students with disabilities as defined by the act were required to have an IEP that specified individual goals. Additionally, student progress regarding these goals was required to be monitored in a continuous fashion to allow for instructional modification as needed (Fuchs, Deno, & Mirkin, 1984). These provisions coincided with the rise of the mainstreaming movement, in which schools were attempting to provide services for special education students in the general education classroom to the maximum extent possible. The addition of special education students to the general education classroom meant that general education teachers

needed to be able to monitor student success on relevant skills and behaviors to determine if the general education classroom was the appropriate setting for these students and to make relevant program recommendations based on outcome data. DBPM was created as a way to monitor progress of special education students towards IEP goals both in special and general education classrooms (Deno & Mirkin, 1977).

According to the training manual (Deno & Mirkin, 1977), DBPM is based on five basic assumptions. First, changes in instructional programming should be considered hypotheses that need to be tested for effectiveness and should not be assumed to be effective without supporting data. Second, time-series designs are the optimal method for assessing effectiveness of such programmatic changes. In order to determine whether a program of instruction is effective for a certain student, the effectiveness of the program must be directly measured or else it will not be clear if it is the program itself that is responsible for behavior change. The third assumption is that special education is an intervention system and thus must be empirically tested. Fourth, in order to use time-series designs to test the effectiveness of special education, it is necessary to determine "vital signs" indicative of educational growth and success (page 14). According to Deno and Mirkin, when DBPM was created, no "vital signs" of educational success had been determined. Using these "vital signs," academic "health" can be defined as the difference between a student's level of performance and the level of performance needed to be successful in that educational environment (page 14). The final assumption central to DBPM is that applying time-series designs to test program effectiveness requires training to ensure accurate interpretation. Data-Based Program Modification (DBPM) is a method of implementing programmatic changes, monitoring student progress through the use of "vital signs", and making educational decisions based on resulting data.

Data-Based Program Modification (DBPM) was not the first progress monitoring technique but was different from other progress monitoring methods being utilized at the time it was created. Data-Based Program Modification (DBPM) uses a long-range goal in addition to short-term objectives. It also provides stringent guidelines for how to create the tests used to measure student achievement, which was different from other programs wherein teachers created their own measures without strict guidance. Further, DBPM provides clear rules for how to use data and determine when instructional change is needed (Fuchs, Deno, & Mirkin, 1984).

As outlined in the program manual (Deno & Mirkin, 1977), DBPM is organized around five decision areas that are each linked to a program phase specifying the activities used to make each decision. An essential feature of DBPM is that each program phase consists of elements of four basic processes: measurement, evaluation, communication/collaboration, and consultation training. In each phase, student progress is directly measured, the data is evaluated, and a team makes a decision as to the best course of action for the student based on the data. Throughout all phases of DBPM, student performance is assessed relevant to typical performance as expected for peers in general education, interventions are based on what works for the individual as determined via frequent progress monitoring, the least restrictive alternative is always considered first, and all decisions are based on data for each individual student.

The first decision area is that of problem selection, which is accomplished through an initial needs assessment in which student performance is directly measured in all relevant areas (i.e. academic disciplines, behavior, etc). Discrepancies between current student performance and desired performance are then evaluated. Following problem selection, a program is chosen through the phase of program planning. In the program planning phase, specific methods for measuring the program are created and details of the proposed program are outlined. This is

when short-term and long-term goals are constructed. The third decision consists of program operationalization and the implementation evaluation phase. In this phase, data-collection and data-use are evaluated to see if the program is being implemented as intended, if a sufficient number of data points are being collected, if graphs are being utilized, if modifications are being made based on outcomes, and if all relevant parties have been a part of the process. DBPM depends on constant progress monitoring of student performance, thus the fourth decision involves program improvement assessed through the phase of progress evaluation. In this phase, the data is analyzed to determine rate of improvement during intervention phases through the analysis of median level, trend, and variability of data points. Each program change should be evaluated compared to the initial assessment and compared to other changes to determine effectiveness. The last decision is that of program certification through an outcome evaluation phase, in which the data is examined to determine the discrepancy between present performance and desired performance. At this point, it is determined if the program has been successful in achieving the original objectives and solving the problem. Through these five decisions and continuous measurement and evaluation, program changes for student progress can be directly monitored for effectiveness and modified as indicated by the data. In the DBPM manual, Deno and Mirkin (1977) provide flow charts and specific details for how to make each decision and conduct each phase to help ensure that correct procedures are being followed.

The effectiveness of DBPM was empirically assessed through a large body of research studies conducted at the University of Minnesota Institute for Research on Learning Disabilities (IRLD; Deno, 2003). A representative study of this research is a group design study by Fuchs, Deno, and Mirkin (1984), in which they compared the effectiveness of DBPM to monitoring as usual on student achievement, student awareness as to their own level of performance, and

teacher perceptions of student progress. Teachers in the DBPM group were trained on the techniques of DBPM. They used these techniques to create specific IEP goals for reading achievement, including an end goal and weekly performance objectives, and to create tests for monitoring of student progress. Teachers in this group monitored progress at least one time per week and graphed the data. If the student did not show adequate progress towards his/her goal after 7-10 measurement points, teachers were instructed to change the reading program for that student. Teachers in the progress monitoring as usual group were free to measure student progress as they chose, which typically consisted of teacher-made tests, workbook exercises, and teacher observations. After 18 weeks of implementation, students of the teachers in the DBPM group showed greater levels of reading achievement than those in the treatment as usual group. Additionally, they were more aware of their own goals and level of progress. Teachers from the DBPM group were more realistic about student progress and response to instructional changes compared to teachers in the control group. Further, teachers in the control group were less specific when asked to describe student goals and progress levels and more uncertain as to student performance. This direct comparison suggested that the DBPM system was effective in increasing student achievement and teacher efficacy in regards to monitoring progress.

As a result of the research on DBPM, the method known as Curriculum-Based Measurement (CBM) was created. As originally conceptualized, the use of DBPM involved the creation of assessment probes to measure students' ipsative and relative performance from a school's individual curriculum. However, this method posed methodological concerns as curriculum varied between schools (Shinn, 2010). In order to address this concern, standardized measurement in the form of CBM was created. Creation of CBM involved delineating the main outcome measures on which to evaluate performance, building measurement systems of relevant

stimuli and scoring techniques to accurately assess those outcome measures, and forming decision rules for determining program effectiveness (Deno, 2003).

Curriculum-Based Measurement

Creation of CBM. When CBM was created, there was a general consensus in education that measuring student progress was important; however, there was little agreement as to the best method for doing so. As explained by Deno (1985), alternate choices included achievement tests and teacher observations. Achievement tests, while providing a good indicator of student performance compared to a normative sample, are often not aligned to curriculum objectives and are not suited for frequent administration or estimates of day-to-day indices of change. Teacher observation of progress, while popular with teachers, has unknown reliability and validity. Deno argues that CBM takes advantages of achievement tests and observation and combines them. The first step in the creation of CBM measures involves observation of curriculum and student performance (Deno & Fuchs, 1987). Peer norms can also be created through use of average student performance on a CBM measure at a certain grade level. With these norms, individual performance can be compared to that of typical student. At the same time, CBM data can be individually-referenced as individual student progress can be compared to itself (Deno, 1985). In this sense, CBM can be both relative and absolute and has elements of both standardized assessments and observations.

According to Deno (1985), the creation of CBM was guided by a number of rules. It had to be reliable and valid, simple and efficient, easy to understand, and inexpensive to use. The creators of CBM intended for it to be responsive to intervention and show growth over time. As part of these qualifications, CBM had to possess adequate reliability, validity, and sensitivity to change, and needed to result in reliable and valid decision-making. Further, these indices needed

to differentiate rates of student performance and group students based on achievement level. Finally, CBM should target easy to measure behaviors in an efficient, cost-effective, and non-intrusive manner.

In terms of determining what and how to measure progress, Deno and Fuchs (1987) distinguish between two choices: performance measurement and progress measurement. Performance measurement involves measuring student behavior on the same task of the same difficulty level over time. Progress measurement involves measuring mastery of curriculum over a period of time using sequences of tasks that a student progresses through. The choice of which type of measurement system to use might depend on the domain being measured. Regardless of the measurement system, a task must be chosen and a type of score. Then, difficulty level (performance) or unit of mastery (progress) must be selected. Following these decisions, the frequency of measurement and mastery criteria must be determined and test samples and procedures must be created. Both types of measurement were considered relevant for CBM.

Part of the initial creation of CBM involved choosing measures that had face validity, meaning that on the surface they seemed as if they would represent progress in a particular area. An example of such measures that were considered for reading CBM included answering reading comprehension questions, filling in missing words in reading passages, defining words embedded in passages, reading aloud from a passage, or reading aloud from a word list (Deno, 1985). However, just because an item has face validity, does not mean that it is a good tool for measuring progress. For example, answering comprehension questions is not something that can always be done quickly and efficiently and might require substantial effort to create multiple probes.

Once behaviors were chosen, the reliability and validity of the measures had to be assessed empirically. Possession of adequate reliability and validity is necessary for any measurement tool, including CBM. Validity of an assessment tool is an estimate of its accuracy to measure what it intends to measure. Additionally, it reflects the degree that the use of tests and the interpretations of resulting scores are appropriate when considering relevant theory and evidence (Gansle et al., 2006; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Two common types of validity relevant for assessment tools are criterion validity and construct validity (McMaster & Espin, 2007). To assess criterion validity, the correlation between the relevant measure and an already well-established measure is examined. The well-established measure should be considered to be an important indicator of performance for that domain (i.e. a standardized assessment) so a high correlation indicates that the assessment is measuring what it intends. Similarly, construct validity can be tested by assessing the degree to which the relevant measure agrees with theoretically-related measures and does not agree with theoretically unrelated measures (AERA, APA, & NCME, 1999). The importance of assessing these types of validity can be highlighted with the reading CBM example mentioned earlier. Some measures that may seem like good indicators of progress do not, in fact, possess adequate criterion validity, such as defining words embedded in passages (Deno, 1985).

Reliability, the degree to which a measure is consistent across implementations, is especially important for CBM given that CBM is used to monitor progress over time to make instructional decisions. Types of reliability that are considered particularly relevant for CBM include alternate-form reliability (reliability across different test forms), test-retest reliability (reliability across occasions), and inter-scorer reliability (reliability across administrator/scorer).

If the measure is not consistently reliable across these dimensions, decisions made based on CBM data may be flawed (Gansle et al., 2006; McMaster & Espin, 2007).

The reliability and validity of CBM measures was investigated thoroughly when CBM was created in the late 1970s and early 1980s at the University of Minnesota. Since that time, a number of studies have been conducted to expand on the results of those foundational studies, enhancing their generalizability and validating the current uses of CBM. The reliability and validity of reading CBM measures in particular have been documented extensively and reading is considered to be the most well-established area of CBM in this regard (McMaster and Espin, 2007; Wayman , Wallace, Wiley, Tichá, & Espin., 2007). A review of reading CBM technical adequacy research was conducted by Wayman and colleagues in 2007. This review focused on reading CBM studies conducted after the foundational studies of the 1970s and 1980s. In a literature search from 1989 to 2007, the authors found 96 studies available for review. Although results varied across studies and across type of CBM assessment, reliability and validity coefficients were generally large ($r > .85$).

Although less established than that of reading CBM, math CBM has a relatively substantive amount of research attesting to its reliability and validity. A review of 32 studies on math CBM conducted in 2007 found moderate support for the technical adequacy of math CBM as currently implemented in schools. Results varied across type of CBM assessment, as well as across grade level, and the authors indicate that further research in this area is needed (Foegen, Jiban, & Deno, 2007). The research base on writing CBM will be discussed in detail later in the paper.

General Characteristics of CBM. Although CBM started as a tenet of special education, the utility of CBM in general education readily became apparent and thus, CBM quickly began to be used for progress monitoring and screening in a wide variety of contexts (Hosp et al., 2007; Shinn, 2010). Besides being a useful technique for monitoring student progress, the recent emphasis on accountability has made CBM even more relevant and important for use in schools today as CBM provides an indicator of which students are at-risk for not achieving grade level standards and thus are in need of intervention (Gansle et al., 2006; McMaster & Espin, 2007). Deno and colleagues (2009) describe how CBM measures can be used to quickly and easily identify students in need of referral within a response to intervention (RTI) model as it requires little effort, is a time-efficient assessment tool, and can be administered in a group context.

Curriculum-based measurement (CBM) has a number of attributes that distinguish it from other measurement systems and that make its use in school environments desirable. A main tenet of CBM is that it is directly aligned to curriculum as it assesses the same content using similar materials and response formats as would be found in a classroom setting (Deno, 2003; Hosp et al., 2007). This alignment means that CBM has a high match between testing and teaching, also known as curricular validity (Deno & Fuchs, 1987). This aspect of CBM is an advantage compared to standardized tests, which often are not aligned to student curriculum, as mentioned previously (Deno, 1985).

Another useful aspect of CBM is that it can be used to assess both specific skills and general outcome measures (GOMs). GOMs are holistic measures based on a combination of skills that provide an overall picture of success in a given area. Just as blood pressure is a GOM used in medicine for general health, oral reading fluency (ORF) is a GOM used in education to

assess overall reading ability (Hosp et al., 2007; Powell-Smith & Shinn, 2004). Similarly, CBM is versatile in that it is intended to be used repeatedly over time in order to provide a measure of a student's progress, but it can also be used for screening purposes to indicate overall level of performance compared to normative criteria. Further, CBM uses standard procedures so that the administration process and scoring for every CBM is identical within a content area.

In order for a progress monitoring technique to be useful within a standards-based accountability system, it needs to have clear standardized methods of scoring, analyzing, reporting, and tracking data. Additionally, there needs to be a way to utilize that information for instructional modification (Quenemoen, et al., 2004). For the most part, CBM meets those criteria as methods for scoring and administration are standardized, norms for typical levels of student performance have been determined, and CBM measures can be easily graphed to monitor progress. The sensitivity of CBM to small changes in student performance and ability to be administered often, combined with the use of graphed data, allows for improved communication of student performance compared to a standardized achievement test. Graphs are easy to read and understand, represent data in a simple manner, and can show growth over time compared to individual goals and/or peer norms (Deno, 1985). In this way, CBM can help improve communication between teachers and parents and between teachers in general education and those in special education (Deno & Fuchs, 1987).

Curriculum-based measurement (CBM) is also an attractive procedure for use in schools in that it is efficient, quick to administer, and requires minimal training for use. Curriculum-based measurement (CBM) may be especially useful for students with learning problems since it is meant to focus on the individual, is sensitive to small changes, and can be administered repeatedly to assess progress (Deno, 2003; Olinghouse & Santangelo, 2010). Marston, Mirkin,

and Deno (1984) found that students referred after 10-weeks of progress monitoring using CBM techniques were significantly more likely to meet learning disability criteria when assessed (80% of the referral sample) versus those who were referred via traditional methods involving teacher opinion (36% of the referral sample). Further, the proportion of boys and the proportion of students with significant behavior problems referred for eligibility evaluations was less when CBM techniques were used compared to traditional methods, suggesting that CBM may help to reduce teacher referral biases.

Another distinguishing feature of CBM is that it requires low inference: conclusions are made directly based on raw-score performance and do not require conversion into percentiles or normal-curve equivalents. Finally, just as with any good assessment instrument, CBM is technically adequate and has established reliability and validity, particularly in the domains of reading and math as previously mentioned (Deno, 2003; Hosp et al., 2007). Both of these last two features highlight the advantage of CBM over the use of teacher observation of progress, which tends to be high inference and has unknown reliability and validity (Deno, 1985).

Curriculum-Based Measurement for Written Expression

As the focus of the current study is on the use of writing CBM, the, creation, use and technical adequacy of such measures will be examined in more depth than that of their reading and math counterparts.

Creation of Writing CBM. The assessment of writing through CBM presents a number of unique challenges. Writing is a complex task that involves a number of inter-related skills, including, but not limited to: idea generation, organization of thought, fluency, vocabulary, syntactic maturity, and use of language conventions (Gansle, Noell, VenDerHeyden, Naquin, & Slider, 2002; McMaster & Espin, 2007). Ideally, writing assessments would be able to

accurately capture all of these unique aspects in a reliable way across multiple populations, environments, and age-ranges.

In creating the writing CBM, the main task was determining which behaviors to measure that would be valid indicators of writing ability and would also be sensitive to change. As determined by face validity and past research, the task for writing CBM that was chosen was the production of a writing sample. Research through the Institute for Research on Learning Disabilities (IRLD) determined that the length of the sample could vary (studies using between 3 and 10 minutes) without affecting the quality of scoring indicators (McMaster & Espin, 2007). In order to determine the relevant behavior to measure, a number of behaviors were considered in the original creation of writing CBM. Deno, Marston, and Mirkin (1982) investigated six different methods of scoring (which measured the relevant behaviors): mean T-unit length (an estimate of grammatical level), total number of "mature words", total number of words written (TWW), word length, words spelled correctly (WSC) and letter sequences written correctly. The authors examined the correlation of these measures with already established criteria, differences in performance across age level, and ability of the measures to discriminate those with learning disabilities from those in general education. Results of this initial study indicated that TWW, WSC, correct letter sequences, and mature words were consistent and strong behavioral representations of writing ability. The validation process for these behaviors, has in fact, been ongoing and as discussed later, some of the original scoring methods have since been brought into question, while additional scoring methods have been developed. Regardless, this example shows the process through which writing CBM was initially created.

Characteristics of Writing CBM. The writing CBM task for upper elementary-aged students has remained consistent since its development and involves the implementation of a 3 through 5 minute writing probe in which the student is given an age-appropriate story starter consisting of a short sentence meant to initiate the writing process (Hosp et al., 2007; Powell-Smith & Shinn, 2004). A common practice is to administer three probes at one time and take the median score; however, single probe administration can be conducted as well (Hosp et al., 2007). Although it is suggested that writing CBM can be used for progress monitoring purposes as a general outcome measure or to pinpoint specific writing problem areas (Olinghouse & Santangelo, 2010), an area that has still not been resolved is that of how to appropriately score such measures. A variety of scoring indicators attempting to assess both quantity and quality of the writing process have been proposed and tested. However, more research is needed to confirm the technical adequacy of many of these measures (Gansle et al., 2002; Hosp et al., 2007). Typically, the same scoring indicator is used over the course of the entire progress monitoring process for an individual student/school (Olinghouse & Santangelo, 2010) so research providing insight and guidance into which scoring method to use is needed. Additionally, it would be problematic if different scoring measures differentially indicate success for the same student since writing CBM is intended to be used to inform treatment and make important relative and absolute instructional decisions. And, as previously mentioned, progress monitoring techniques within a standards-based system should have clearly defined scoring procedures (Quenemoen et al, 2004), which is currently lacking for writing CBM.

Although many current researchers in the field have doubts as to the accuracy and reliability of these measures, the most common scoring methods for writing CBM are total words written (TWW), total number of words spelled correctly (WSC) and number of correct writing

sequences (CWS; Hosp et al., 2007; Powell-Smith & Shinn, 2004). TWW is a measure of the total number of words written by the participant, regardless of spelling or context, and WSC is a frequency measure of the number of words correctly spelled in consideration of context. CWS are "two adjacent writing units (words and punctuation) that are correct within the context of what is written," (Powell-Smith & Shinn, 2004). A number of other scoring measures have been suggested and tested to varying degrees. In the research of Deno, Mirkin, and colleagues at the IRLD, other scoring measures included mean T-unit length, number of large words (containing more than 7 letters), number of mature words (based on a standardized frequency index of word usage), correct letter sequences (CLS), words spelled incorrectly (WSI) and total number of letters (Deno et al., 1982; Deno, Mikrin, & Marston, 1980; & Tindal, Germann, & Deno, 1983). Since then, Gansle and colleagues (2002; 2004; 2006) have investigated a number of different measures, including, but not limited to, number of different parts of speech, total punctuation, correct punctuation, correct capitalization, words in complete sentences, various computer-scored variables, and complete, simple, or fragmented sentences. Correct minus incorrect writing sequences has also been examined (Henderson, 2009; McMaster & Campbell, 2008)

Institute for Research on Learning Disabilities (IRLD) Studies. Writing CBM measures were first developed and investigated by Deno and Mirkin as part of the original DBPM/CBM research of the IRLD. As reported in a review of 28 studies on writing CBM by McMaster and Espin (2007), the foundational research found that these measures had high technical adequacy overall and were acceptable for use as progress monitoring measures. All of the IRLD studies evaluated writing CBM using elementary-aged children. Populations tested included students in general education, students in special education, low achievers, and children with diagnosed learning disabilities (LDs). A review of these studies follows. It is worth noting

that there is not a clear standard for what constitutes adequate reliability and validity of measures. It has been suggested that for CBM coefficients of $r > .80$ are strong, those between .70 and .80 are moderately strong, r values of .60 to .70 are moderate, and r values falling below .60 are weak (McMaster & Espin, 2007). Other studies on CBM have used similar but slightly different criteria, with r coefficients > .70 considered strong, r values between .50 and .70 considered moderate, and r values below .50 considered weak (Wayman et al., 2007). Results should be interpreted relative to each other and in light of these past criteria.

Based on prior research, Deno, Mirkin, and Marston (1980) evaluated the criterion validity of five different scoring procedures with three different empirically-validated standardized measures of writing: the Test of Written Language (TOWL; Hammill & Larsen, 1978), the written portion of the Stanford Achievement Test (SAT; Madden et al, 1978), and the Developmental Scoring System (DSS; Lee and Canter, 1971). The sample included students in general education and those diagnosed with LD in grades 3-6. In the first study, the authors found moderate-high correlations with the raw total of the TOWL for number of large words (r = .56-.66), number of mature words (r = .65-.78), total words written (TWW; r = .63-.82), and words spelled correctly (WSC; r = .63-.88). In Study 2, a similar pattern of results was found for the TOWL. For the written portion of the SAT, correlation coefficients were r = .42-.72 for large words, r = .52-.72 for mature words, r = .56-.71 for TWW, and r = .65-.77 for WSC. The final study assessed the criterion validity of the various scoring procedures with the DSS, finding low-moderate correlations for large words (r = .23-.35), and moderate-high correlations for mature words (r = .54-.74), TWW (r = .65-.88), WSC (r = .67-.84), and correct letter sequences (CLS; r = .64-.86). Other important findings of this study included that there were no significant differences in scores when using a 3-, 4-, or 5-minute probe and no significant differences when

using a pictorial prompt, an oral story starter, or written topic sentences as the writing probe. When combining results across studies, the authors concluded that TWW, WSC, and mature words were the best scoring methods as they had the least differences across administration formats and had consistent validity scores. Additionally, it was found that the writing probes reliably distinguished those with a diagnosis of LD from those students in general education.

Videen, Deno, and Marston (1982) conducted a similar study in order to assess the criterion validity of correct writing sequences (CWS) as a scoring procedure. General education students in grades 3-6 were provided story prompts. Criterion validity was found for CWS with the DSS (r = .49), the TOWL (r = .69), a holistic rating of impression of writing quality (r = .85), TWW (r = .91), and WSC (r = .92). Based on the results of the study, the authors concluded that CWS was a valid scoring system for writing CBM.

In order to assess reliability of the writing CBM probes, Marston and Deno (1981) conducted a follow-up consisting of four studies assessing various forms of reliability for four different scoring procedures. The authors administered writing probes to students with and without a learning disability in grades 1-6. In order to assess test-retest reliability, writing probes were given on the same day and again three weeks later. Reliability coefficients for mature words were r = .57 for probes given within the same day and r = .50 for probes given three weeks later. For TWW the coefficients were r = .91 and r = .64 for one day and three weeks respectively. Respective coefficients of r = .81 and r = .62 were obtained for WSC for one day and three weeks. And finally, coefficients for CLS were r = .92 for one day and r = .70 for three weeks. Additionally, the parallel-form reliability was assessed between probes administered through pictorial prompts, verbal story starters, and written topic sentences. Parallel-form reliability coefficients were r = .74-.79 for mature words, r = .79-.85 for TWW, and r = .81-.87

for WSC. In an analysis of split-half reliability, Cronbach's alpha was calculated for the different scoring measures. The following alphas were obtained: mature words = .74, TWW = .87, WSC = .70, and CLS = .87. For the final reliability study, the authors assessed inter-scorer reliability among the measures and obtained *r* = .90-.94 for mature words and *r* = .98-.99 for TWW, WSC, and CLS. Overall, the authors concluded that reliabilities for all scoring systems were large enough to consider writing CBM technically adequate when added to the validity results of the prior study.

A number of other studies were conducted on the technical adequacy of writing CBM as part of the IRLD initiative. Alternate-form reliability for TWW was tested in both low achieving students and students in general education in grades 1-5 and was found to have reliability coefficients between .51 and .71 (Shinn, Ysseldyke, Deno, & Tindal, 1982). Similar results were found (*r* = .55-.89) for the alternate-form reliability of WSC in a sample of low achieving students in grades 3-6 (Fuchs, Deno, & Marston, 1982). Another study using a sample of fourth and fifth grade general education students found alternate-form reliability of *r* = .71 for TWW and *r* = .70 for CLS (Tindal, Germann, & Deno, 1983). In a study of general education students in grades 1-6, alternate-form reliability was found to be *r* = .73 for TWW, *r* = .72 for WSC, and *r* = .93 for CLS (Tindal, Marston, & Deno, 1983). Further, all of the studies that assessed inter-scorer reliability found it to be above *r* = .90 for all scoring measures, grades, and populations (Deno et al., 1982; Marston, Deno, & Tindal, 1983; Tindal, Marston, & Deno, 1983; Videen, Deno, & Marston, 1982).

Additionally, a handful of studies examined reliability of measures from fall-spring. Tindal, German, and Deno (1983) found fall-spring coefficients of *r* = .56 for TWW and CLS in a sample of fifth grade general education students. Deno and colleagues (1982) examined fall-

spring reliability for TWW, WSC, CLS, and words spelled incorrectly (WSI) in a sample of 1-6 grade general education students. The coefficients for first graders were low (r = .20-.47) but the coefficients for the remaining grades were large (r = .60-.86). Overall, the IRLD research suggested that writing CBM has acceptable reliability and validity for use with elementary school students across grades and populations using a number of different scoring techniques.

Extension of IRLD Research. Empirical research conducted since the foundational research on writing CBM has been less conclusive in regards to its technical adequacy and has brought some of the conclusions of the original research into question, specifically in regards to the most common scoring procedures. Although research has extended past elementary-aged students, for the purposes of this review, only the elementary-aged studies will be included. A review of other studies can be found in McMaster and Espin (2007).

Tindal and Parker (1991) examined the consistency of scoring procedures, criterion validity, improvement across a year, and ability to discriminate between general and special education students of writing CBM in a sample of third-fifth grade students. The main purpose of this study was to validate the use of these measures across a full range of students, including those with LD, those deemed low achieving based on standardized test performance, those deemed low achieving by general education teachers, those of average general education performance, and those receiving special education services. Results indicated that TWW, WSC, and CWS had correlations above r = .85 with each other, but had low-moderate correlations with qualitative measures based on teacher rating. Additionally, TWW correlated at r = .22 with the SAT, while WSC correlated at r = .28 and CWS at r =.41. Using a principle components analysis, 81% of the variance was accounted for by three factors: the first including TWW, WSC, CWS, and total number of word sequences, the second including percent of CWS, and

ratings of mechanics usage and conventions, and the third including ratings of story idea and cohesion/organization. Taken together, these results suggest that qualitative measures might be needed in addition to the quantitative measures of TWW, WSC, and CWS to explain student performance on writing CBMs. Additionally, most measures were able to capture student improvement across the year, although the results were not consistent across measures. The different scoring methods were all able to discriminate between students with learning disabilities and those in general education but were not as effective at distinguishing between low achievement and average achievement in general education. The authors concluded that more research would be required in order to determine the best scoring procedures for discriminating between groups and measuring growth in special education.

Parker, Tindal, and Hasbrouk (1991) conducted a study with a large sample of students in grades 2-5 in both general and special education in order to determine the sensitivity of writing CBM for special education screening and for differentiating among different levels of achievement. The authors used histograms with imposed normal curves, percentile graphs, and standard error of measurement (SEM) bands to examine the sensitivity of TWW, WSC, CWS, percent of WSC, and percent of CWS. They found that percent of WSC had the greatest measurement sensitivity, followed by percent CWS when scores from second grade were excluded from the sample. However, the scoring measures all had large SEMs and did not effectively distinguish students at the lower ends of the scale, suggesting that writing CBM measures may possess low levels of sensitivity for students with lower scores, which is problematic when trying to distinguish those who may be at-risk for writing failure (typically in the bottom 25th percentile). Additionally, the authors examined the criterion validity of the different scoring procedures using a holistic rating of writing effectiveness (rated on a scale from

1-7) and found correlations of r = .36-.49 for TWW, r = .54-.64 for WSC, r = .58-.61 for CWS, r = .48-.67 for percent WSC, and r = .43-.70 for percent CWS. These measures are lower than the correlations found by Videen, Deno, and Marston (1982) of r = .85 for holistic ratings, suggesting that either the scoring measures or the criterion measure may be inconsistent.

Gansle, Noell, Vanderheyden, Naquin, and Slider (2002) investigated the alternate-form reliability, inter-scorer reliability, and criterion validity of 19 different scoring measures for writing CBM of third and fourth grade students. The measures were examined through correlation coefficients and entry in a multiple regression. Criterion validity was examined in regards to scores on the *Louisiana Educational Assessment Program* (LEAP), the *Iowa Test of Basic Skills* (ITBS), and teacher rankings. Refer to their text for specific details on all 19 measures and complete results. Based on combined results, the authors concluded that CWS sequences was a valid indicator of writing performance, that TWW may not be as technically sound as prior research suggested, and that correct punctuation might be a useful scoring measure.

Malecki and Jewell (2003) examined the technical adequacy of writing CBM in students in grades 1-8 in the fall and the spring. They used a MANOVA to analyze results for a composite of dependent measures consisting of TWW, WSC, CWS, correct minus incorrect writing sequences (CIWS), percentage of WSC, and percentage of CWS. The results indicated that significant differences between scoring measures were present across all grades. Additionally, scores were significantly higher in the spring compared to the fall, suggesting that the scoring variables exhibit growth over time.

Gansle et al. (2004) conducted a study in order to assess the criterion validity of writing CBM samples for third and fourth grade students with the *Woodcock Johnson—Revised* (WJ—

R; Woodcock & Johnson, 1989) Writing Samples subtest. They examined six different scoring procedures: TWW, total punctuation marks, correct punctuation, words in complete sentences, CWS, and simple sentences. Correlation coefficients with the WJ-R were as follows: r = .23 for TWW, r = .42 for total punctuation marks, r = .34 for correct punctuation, r = .35 for words in complete sentences, r = .36 for CWS, and r = -.05 for simple sentences. When these variables were entered into a multiple regression equation, 43% of the variance was predicted using total punctuation marks (β = .62), simple sentences (β = -.55), and words in complete sentences (β = .39). Interestingly, TWW was not a significant predictor in the regression and CWS was no longer significant when other variables were included.

A study by Jewell & Malecki (2005) investigating the validity of writing CBM measures to predict scores on the SAT, an analytic scoring system, and English-language arts grades across second, fourth, and sixth grade students found results suggesting that simple fluency measures, such as TWW, WSC, and CWS may become less valid as students progress in school. The predictive validity of these measures decreased as students got older. However, percent WSC, percent CWS, and CIWS had higher predictive validity across grades. Additionally, TWW did not correlate significantly with percent WSC or percent CWS, suggesting that longer writing samples were not necessarily reflective of more accurate writing. The authors concluded that CIWS is a promising indicator of both fluency and accuracy and correlates highly with criterion measures.

In a similar study to Jewell and Malecki (2005), Wessenburger and Espin (2005) examined alternate-form reliability and criterion validity of TWW, CWS, and CIWS across the fourth, eighth, and tenth grades. Alternate-form reliability for all measures in the fourth grade was high with *r*s above .80 for TWW and CWS and *r*s above .70 for CIWS. However, alternate-

form reliability coefficients decreased as the students got older. A similar pattern was found for criterion validity with the *Wisconsin Knowledge and Concepts Examination* language arts portion. For fourth graders, criterion validity of TWW was r = .36-.45, for CWS it was r = .56-.62, and for CIWS it was r = .67-.68. These coefficients also decreased with age. Results of these two studies suggest that CIWS is a technically adequate measure of writing for students in upper elementary school and that writing CBM as traditionally utilized may not be valid beyond the elementary grade levels.

To expand on previous research, Gansle, Vanderheyden, Noell, Resetar, and Williams (2006) conducted a study using a large sample of students in grades 1-5 to examine the criterion validity of seven different scoring measures with the *Stanford Achievement Test, Ninth Edition* (Stanford-9; Harcourt Brace Educational Measurement, 1996) writing section. Correlations with the total score of the Standford-9 written portion were as follows: r = .34 for TWW, r = .43 for CWS, r = .38 for WSC, r = .39 for correct punctuation, r = .28 for correct capitalization, r = .36 for complete sentences, and r = .41 for words in complete sentences. Further, TWW, CWS, and WSC formed a highly correlated cluster and correct punctuation, complete sentences, and words in complete sentences formed a moderately correlated cluster. The authors suggest that these two distinct clusters might represent different aspects of writing and should both be considered when scoring writing CBM. Additionally, the authors examined the test-retest reliability of the measures. Test-retest reliability for TWW, CWS, and WSC was all above r = .70 as consistent with prior research. Test-retest reliability for correct punctuation was .64, for correct capitalization it was .44, for complete sentences r equaled .65, and for words in complete sentences r was equal to .61. Overall, the results suggest that correct capitalization may not possess sufficient technical adequacy.

McMaster and Campbell (2008) examined the alternate-form reliability and criterion validity with the TOWL for scoring procedures of TWW, WSC, CWS, and correct minus incorrect writing sequences (CIWS) across fall and spring for narrative prompts, pictorial prompts, and expository prompts in the third, fifth, and seventh grades. The authors found considerable variability across grades. However, across grades, CWS and CIWS used with narrative prompts were the most consistently reliable and valid.

In an unpublished dissertation, through the use of Principal Components Analysis (PCA), Henderson (2009) found further support for the clusters identified by Gansle et al (2006). Additionally, Henderson looked at the predictor-criterion relationship between a number of scoring methods for writing CBM in the fall, winter, and spring in predicting performance of third grade students on the *integrated Louisiana Educational Assessment Program* (*i*LEAP) English, Language Arts section. Using multiple regression, Henderson found the best predictors of student performance to be WSC in the fall, number of complete sentences in the winter, and percent of CWS in the spring. Additionally, CWS had the highest discrimination in terms of sensitivity, specificity, positive predictive power, and negative predictive power.

Other research has been conducted examining the technical adequacy of writing CBM in older students but this research will not be summarized here as the focus of the current study is on the upper elementary level. Additionally, there is a growing body of research around the use of different forms of writing CBM for students in younger grades, specifically grades K-2. These measures include alternate procedures, such as sentence copying, word copying, letter prompts, picture-word sentence prompts, picture theme prompts (McMaster, Du, & Pétursdóttir, 2009; McMaster & Campbell, 2008; Parker, McMaster, Medhanie, & Silberglitt, 2011), writing two sentences from a prompt (Coker & Ritchey, 2010), sentence dictation, and word dictation

(Lembke, Deno, & Hall, 2003). The research on newer forms of CBM for early writers will not be reviewed here since this study will examine traditional measures of writing CBM.

Overall, results of the research on traditional writing CBM at the elementary level indicate that more research is needed at this level in regards to test-retest reliability, alternate-form reliability and the lowered criterion validity measures obtained in more recent studies compared to the original studies (McMaster & Espin, 2007). The research in this area remains inconclusive as different scoring measures have performed differently across studies and across administration periods (i.e. fall, winter, spring). Additionally, as reported in the introduction to a special issue regarding the use of CBM within a standards-based system, research on CBM in this domain is lacking. The Office of Special Education Programs (OSEP) funded the Research Institute on Progress Monitoring (RIPM) in order to promote further study of this area. Specifically, the goal of the research is to develop methods of progress monitoring that can be used across environments, age groups, skill levels, and curricula (Wallace, Espin, McMaster, Deno, & Foegen, 2007). As described in the review above, writing CBM lacks these qualities and thus merits further study.

Generalizability Theory

An advanced statistical approach to studying technical adequacy that could add to the research base and help to fill in knowledge gaps for writing CBM is Generalizability (G) theory. G theory was created by Cronbach, Gleser, Nanda, and Rajartna, (1972) as a way to assess the reliability of behavioral measures in a manner that accounts for some of the disadvantages of classical test theory. An implicit assumption of classical test theory is that a person's true score is constant and that any difference in observed scores must be due to measurement error. Classical

test theory further assumes that there are only two sources of variance: that explained by the true score and that explained by error. An observed score can be expressed as:

$$X = T + E, \quad (1)$$

where X is the observed score, T is the true score, and E is the error score (Brennan, 2011). The error term is thought to be random and can result from a number of different sources, which are not parsed out individually. As a result of these assumptions, reliability in classical test theory is examined with no reference to context (Brennan, 1992; Hintze, Owen, Shapiro, & Daly, 2000; Webb, Shavelson, & Haertel, 2006). Reliability in classical test theory is typically expressed as a coefficient ranging from 0-1.0 and can be interpreted as the proportion of variance accounted for by the true score. However, there are a number of ways to measure reliability in classical test theory, thus any one test can have a large number of reliability coefficients (i.e. alternate form reliability, internal consistency, test-retest reliability, etc.). These different reliability coefficients can sometimes result in very different estimates of reliability for a single measure (Webb, et al., 2006).

In contrast to classical test theory, G theory allows multiple sources of variance to be assessed, in addition to the true score. Instead of assuming one random error term, G theory assumes that there are many sources of error that contribute to overall variance and these sources are thought to be systematic, as opposed to random. An observed score can be expressed as:

$$X = \mu_p + E_1 + E_2 + \ldots E_{n,} \quad (2)$$

where X is the observed score, μ_p equals the universe score, E_1 represents error from the first source, E_2 represents error from the second source, and so on for each source measured in the study (Brennan, 2011). This conceptualization means that these sources can, in fact, be measured individually and accounted for separately (Brennan, 1992; Webb et al., 2006). A universe score

(μ_p) is the expected value of observed scores over all possible measurement occasions, with each occasion consisting of a different random sample of conditions specified in the study. In this sense, G theory assumes that any single observation (or measurement) is representative of all possible observations in that context (Brennan, 2011). In G theory, reliability (how consistent observations are) and validity (how accurate the sampled observations are) for a certain measurement system for a particular subject are estimated simultaneously, rather than separately (Hintze, 2005). Instead of referring to the reliability or validity of measures, G theory refers to the dependability of measures, which encompasses both concepts (Hintze & Matthews, 2004). Dependability is an estimate of the accuracy of generalizing from the observed score to the average score a person would have for all possible testing occasions under the conditions included in the study (Shavelson & Webb, 1991).

As alluded to in the preceding paragraph, G theory examines reliability and variance within the context of the particular assessment environment. This context is termed a universe of admissible observations (Brennan, 1992; Shavelson & Webb, 1991; Webb et al., 2006). The universe in any particular G theory study is determined by the researcher and consists of any observations that the researcher would consider interchangeable for the purposes of the decision at hand (Shavelson & Webb, 1991). For example, two different researchers may wish to examine the error due to rater for a particular assessment. However, one researcher may want to look at those with an advanced degree as raters and the other researcher may wish to look at undergraduate students as raters. Although both researchers may be examining the same variable (i.e. rater) for the same assessment, they have defined the universe differently. In this sense, it is necessary to fully understand the universe being investigated in a G theory study before generalizing the results (Brennan, 1992).

Sources of error in G theory are referred to as facets. Some common examples of facets include occasion, form, rater, setting, dimension, and scoring method (Hintze et al., 2000). Although person is sometimes referred to as a facet, for the most part, person is the object of measurement (Shavelson & Webb, 1991). Facets can be crossed or nested. When facets are crossed, each participant receives each condition of one facet combined with each condition of another facet. For example, in an investigation of the facets of items and occasion, each item would be presented on each occasion. If the design is fully crossed, then all conditions of each facet appear with all conditions of every other facet. In a nested design, two or more conditions of the nested facet are combined with only one condition of another facet (Shavelson & Webb, 1991; Webb et al., 2006). For example, a different test form may be used for each testing occasion so the form facet would be nested within the occasion facet.

Facets can also be random or fixed. A random facet is one in which members of that facet have been randomly sampled from the larger universe of all possible members. The chosen sample is much smaller than the entire universe population and is considered to be interchangeable with any other same-sized sample randomly chosen from the universe. With fixed facets, the entire universe of concern is represented in the study. In this case, the researcher either does not wish to generalize beyond the conditions represented in that particular study or the universe is entirely represented in the study (Shavelson & Webb, 1991; Webb et al., 2006).

G theory technique is analogous to repeated measures ANOVA, where each possible source of error is tested for its contribution to the overall variance. Variance due to main effects of facets and that due to interactions between facets is examined. For example, with a two facet model investigating the effects of rater (r) and time (t),

$$\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(tr) + \sigma^2(ptr), \quad (3)$$

variance of the observed score in the context of rater and test (the left side of the equation) is equal to the independent variance for person and each facet plus the variance attributable to all possible interactions among person, rater, and test (the right side of the equation; Brennan, 2011). G theory also provides for the acquisition of a G coefficient, which is an estimate of the overall variance explained by the entire model, specifically in regards to the proportion of variance due to person (i.e. the individual). The coefficient is obtained using the variance components from equation (3) to obtain an estimate of variance due to person and an estimate of variance due to error. The ratio of person to error variance is used to determine the coefficient. The equations used to calculate G coefficients are presented in the section on Decision Theory. A G coefficient is similar to a reliability coefficient in classical test theory and can be interpreted using the same metric (i.e. .80 and above is considered a very good coefficient). It is an approximation of the variation among individual scores that is systematic and not the result of error (Brennan, 1992; Hintze, et al., 2000; Lai et al., 2012; Shavelson & Webb, 1991; Webb et al., 2006).

Decision Studies. A decision study (D study) is typically conducted as part of a G study. G studies investigate variance for single combinations of facets, whereby D studies look at average scores across facets. The purpose of a D study is to use the results of the G study to determine optimal measurement systems involving the facets at hand. Just as a universe of admissible observations is specified in a G study, a universe of generalization is specified in a D study. A universe of generalization describes the facets involved in the measurement system. For example, a person may wish to determine how many of a particular assessment probe rated by a certain rater are needed in order to obtain a reliable estimate of subject performance. As previously mentioned, this process involves looking at a person's average score across relevant

facets. In an instance where the facets are specified to a certain population, such as in the previous example, the universe of generalization is considered to be fixed. Continuing with this example, the universe of generalization in a D study could also involve investigation of a measurement procedure when used by a random sample of raters or probes. In this case the universe of generalization would be infinite, or random. D theory studies allow you to estimate the variance attributable to person versus that attributable to other facets for any conceivable number of facet combinations. For example, you could estimate the reliability of a particular measurement system for two different raters and compare that to the reliability obtained when using three different raters, etc. This technique allows you to determine the combination of facets that allows for the most reliable measurement results.

Using decision theory, there are two types of G coefficients that can be calculated: absolute and relative. Absolute G coefficients are used when the interest is in making within-subjects decisions and relative G coefficients are used when decisions are being made between individuals. Absolute coefficients reflect the degree of dependability for a measure in reflecting individual performance. With absolute decisions all variance components aside from the object of measurement (i.e. person) contribute to measurement error. Continuing from the previous example when using a facet of rater and time, the absolute error variance ($\sigma^2(\Delta)$)can be expressed as,

$$\sigma^2(\Delta) = \sigma^2(t)/n'_t + \sigma^2(r)/n'_r + \sigma^2(pt)/n'_t + \sigma^2(pr)/n'_r + \sigma^2(tr)/(n'_t n'_r) + \sigma^2(ptr)/(n'_t n'_r) \quad (4)$$

with variance components corresponding to those in equation (3) for each facet and interaction divided by the number of that facet being considered (i.e. n'_t is the number of times being considered for that equation; if the researcher wishes to consider dependability for 2 different measurement times, the value would be equal to two). Following determination of absolute

error variance, an absolute G coefficient, sometimes referred to as a D coefficient (Φ) can be obtained using the following equation,

$$\Phi = \sigma^2(p)/ [\sigma^2(p) + \sigma^2(\Delta)], \quad (5)$$

obtaining a ratio of the variance due to person relative to that due to person plus error.

Relative coefficients reflect the degree of dependability of a measure to compare individuals, examining the effects of various sources of error on an individual's ranking within a group. With relative decisions, only sources of error that reflect interactions between facets and the object of measurement (i.e. person) contribute to measurement error (Hintze et al., 2000; Shavelson & Webb, 1991; Webb et al., 2006). For our example, relative error variance would be expressed as,

$$\sigma^2(\delta) = \sigma^2(pt)/ n'_t + \sigma^2(pr)/ n'_r + \sigma^2(ptr)/ (n'_t n'_r), \quad (6)$$

with a relative G coefficient ($E\rho^2$) of,

$$E\rho^2 = \sigma^2(p)/ [\sigma^2(p) + \sigma^2(\delta)]. \quad (7)$$

Using these equations, the researcher can choose a level of reliability they wish to investigate and calculate varying combinations of facets to obtain that number (Shavelson & Webb, 1991; Brennan, 1992). If a researcher wishes to determine conditions that result in a reliability of .80 for screening purposes (i.e. low-stakes decisions) or those that result in a reliability of .90 for diagnostic use (i.e. high-stakes decisions), he or she can do so.

D studies take the information from a G theory study and use that information to advise practical application of the measurement system at hand for a desired purpose. For instance, following a G theory study investigating direct observation techniques of on-task behavior for 14 students, two times a day, for 10 days, Hintze and Matthews (2004) conducted a D study for the same facets. Results of the D study indicated that in order to obtain reliable samples of on-task

behavior (using reliability above .80), an individual student would need to be observed four times a day over 40 days. The same concepts of facets, crossed and nested designs, and absolute and relative decisions apply to D studies as they do to G studies (Webb et al., 2006).

Multivariate G Theory. If the dependent measure of interest does not consist of a single score but rather a composite of scores, multivariate G theory can be used. Rather than investigating each subscale or measure independently in separate G theory analyses, the composite score can be investigated in one multivariate G theory analysis. Multivariate G theory uses the same logic and statistical techniques as univariate G theory but instead of decomposing scores into variances, it decomposes scores into matrices of variance and covariance for universe scores and sources of error (Webb & Shavelson, 1981). In addition to obtaining estimates of variance from expected mean squares, as is the process with an ANOVA, expected mean products, reflecting covariance among facets, are used (Webb, Shavelson, & Maddahian, 1983). A multivariate G coefficient is representative of the ratio of composite universe score variance to composite total score variance (composite universe score variance plus composite error variance; Brennan, 2010). Joe and Woodward (1976) developed a multivariate G coefficient using the following equation,

$$\mathcal{P}^{2=} \frac{\alpha' V_p \alpha}{\alpha' V_p \alpha + \alpha' V_i \alpha / n'_i + \alpha' V_j \alpha / n'_j + \alpha' V_{pi} \alpha / n'_i + \alpha' V_{pj} \alpha / n'_j + \alpha' V_{ij} \alpha / n'_i n'_j + \alpha' V_e \alpha / n'_i n'_j} \quad (8)$$

where V equals a matrix of variance and covariance components estimated from mean square matrices, n'_i and n'_j are equal to the number of conditions of facets i and j in a D study and α is equal to a vector of canonical coefficients that maximizes the ratio of universe-score variation to universe-plus-error score variation. To obtain estimates of $p^2$and α, the following equation can be used,

$$[V_p - p_s^2(V_p + V_\Delta)]\alpha_s = 0. \quad (9)$$

V_Δ is the multivariate equivalent to $\sigma^2(\Delta)$ and s refers to the r roots ($s = 1,\ldots,r$) of (9). Equations (8) and (9) refer to equations for making absolute decisions (Webb & Shavelson, 1981). Although not specified as such, it logically follows that the equation could be modified for relative decisions by only including variance and covariance components due to interactions with person and other facets. This equation would look as such,

$$\textit{relative g coefficient} = \frac{\alpha' V_p \alpha}{\alpha' V_p \alpha + \alpha' V_{pi} \alpha / n'_{i+} + \alpha' V_{pj} \alpha / n'_j + \alpha' V_e \alpha / n'_i n'_j,} \quad (10)$$

with V_δ replacing V_Δ in equation (9) as the multivariate equivalent to $\sigma^2(\delta)$.

Additionally, a multivariate G theory analysis can be followed up with D studies on each measure considered independently (Brennan, 2011), similar to the process of following up a multivariate ANOVA with a series of univariate ANOVAs.

Generalizability and Decision Studies on CBM. G theory has been used to examine the technical adequacy of CBM for both reading and math. Hintze, Owen, Shapiro, and Daly (2000) applied G theory techniques to a previously collected data set of reading CBM probes. Their sample consisted of 160 general education students from grades 2-5 who were given two one-minute reading probes twice a week over an eight week assessment period. One of the two weekly probes consisted of a passage from a literature-based program and the other weekly probe was a passage from a skills-based program. The outcome measure analyzed was number of words read per minute for each CBM probe. For the G study, the authors used a repeated measures ANOVA to examine the variance due to person, grade, method (different type of CBM), and occasion, finding that most variance explained was due to person, followed by grade. The authors suggest that based on these findings, reading CBM is a reliable method for distinguishing between student performance and grade. As part of the D study the authors calculated the absolute G coefficient and obtained a G_{abs} of .90. This can be interpreted as

indicating that CBM implemented two times per week in grades 2-5 over an eight week period produced highly dependable results for making intra-individual decisions. The authors also examined the dependability of using only one source for the CBM passages and found a G_{abs} equal to .82, suggesting that dependability was still high even when using one source of material for the probes. Additionally, they calculated a relative G coefficient and obtained a G_{rel} equal to .99, indicating that the parameters of the study were also highly dependable for making inter-individual decisions. The G_{rel} obtained when using only one CBM source over only four weeks was equal to .98, also indicating very high dependability for making inter-individual decisions using only one source material and half of the initial measurement period. Further, using only three passages produced a G_{rel} of .95.

In a second study, Hintze, Owen, Shapiro, and Daly (2000), used a similar procedure to examine the dependability of reading CBM for measuring growth rates when using instructional level probes versus challenging level probes. Eighty students in first through fourth grade were given two reading CBM probes twice per week for 10 weeks. One probe consisted of a passage from the instructional level for the grade of the student being tested and one probe consisted of a passage from the challenging level. The number of words read per minute on each CBM probe was the score used as the outcome in data analysis. Similar to the first study, a repeated-measures ANOVA was conducted for the G study, indicating that person and grade contributed the most to variance. In the D study, a G_{abs} of .80 was obtained, suggesting that dependability is sufficient for making intra-individual decisions using instructional level and difficult level CBM probes administered twice a week for 10 weeks. Further, a G_{abs} of .67 was obtained when investigating only one set of probes across only two grade levels. The authors interpreted these results to mean that CBM probes are less dependable for making intra-individual decisions when

only one source of material is used (either an easy source or a difficult one). For decisions involving inter-individual comparisons using the original measurement procedure, a G_{rel} of .98 was obtained. When using only five weeks of measurement the G_{rel} was .96 and the G_{rel} over only three weeks of measurement was .95. Using just three passages resulted in a G_{rel} of .88. These results suggest that a number of combinations were dependable for making inter-individual decisions using reading CBM in this manner.

Hintze and Pelle Petitte (2001) used G theory to examine the dependability of reading CBM across general and special education. Twelve students in a third and fourth grade classroom were administered one of 16 reading CBM probes (using instructional level passages) two times a week for eight weeks. Each probe was only given to each student one time. Six students were receiving special education services and six students were not. Number of words read per minute was the outcome measure utilized in analysis. A repeated-measures ANOVA was used for the G study, indicating that person accounted for the most variance, followed by group (general versus special education). In the D study, a G_{abs} of .88 and a G_{rel} of .99 were obtained. These results suggest that using two weekly CBM probes over the course of eight weeks produced dependably results for both intra- and inter-individual decisions for students in general and special education considered separately. Further, the authors analyzed the results obtained when collapsing students into one group (i.e. combining the general and special education groups) and obtained a G_{abs} of .80 and a G_{rel} of .98. The authors interpreted these results to mean that reading CBMs are highly dependable when making both intra- and inter-individual decisions for students across general and special education.

Mercer and colleagues (2012) examined the generalizability of Maze CBM, a type of reading CBM that assesses the ability of students to provide missing words in passages, for 272

students in the third through fifth grade. Nine probes were administered to each student over the course of three days and number of correct choices was examined for one-minute, two-minute, and three-minute probe lengths. Reliability was found to increase as probe length increased and was higher overall for fifth grade. D study analyses revealed that for making inter-individual decisions with students in the third and fourth grade, three 3-minute probes would be needed to obtain reliability greater than .80 and in fifth grade, two 3-minute probes would be needed. For high-stakes decisions involving reliabilities greater than or equal to .90, five 3-minute probes would be required in the third and fourth grades and three 3-minute probes in the fifth grade. D studies examining intra-individual decisions, indicated that to obtain reliabilities greater than .80, three probes were needed in third grade, four in the fourth grade, and two in the fifth grade. In order to obtain reliabilities greater than .90, more than five probes would be required in the third and fourth grade and four probes in the fifth grade.

The generalizability of easyCBM® reading assessments was examined in a series of technical reports by Lai, Park, Anderson, Alonzo, and Tindal (2012). Students in grades 1-5 were administered three or four different CBM probes across two separate testing sessions. Across a number of separate analyses by form and grade, person accounted for the majority of the variance. A D study revealed reliabilities greater than .80 for using only one form on one occasion.

In a study using G theory to examine math CBM by Hintze, Christ, and Keller (2002), 67 students in grades 1-5 were given three single-skill and three multiple-skill math CBM probes in order to examine variance due to person, grade, type of probe, and probe form (i.e. comparing three different forms within one type of probe). The outcome measure was the number of correct digits per minute. The authors found that approximately half of the variance was explained by

differences among participants and grade level. However, the type of probe used accounted for close to 20% of the variance, suggesting that interpretation of performance on one type of probe should not be generalized to how a student would perform on other types of probes and across other skills. For single-skill probes considered alone, a G_{abs} of .96 and a G_{rel} of .98 were obtained. For multiple-skill probes, a G_{abs} of .95 and a G_{rel} of .75 were obtained. These results suggest that single-skill probes are highly dependable for making both intra- and inter-individual decisions and that multiple-skill probes are dependable for intra-individual decisions but are less dependable for inter-individual decisions.

Christ, Johnson-Gros, and Hintze (2005) examined the generalizability of math CBM probes for 104 fourth and fifth grade students across durations of 1, 2, 3, 4, 5, and 6 minutes. Inter-individual differences and test duration combined accounted for less than half of the overall variance, suggesting that measurement error contributed highly to the variance in this model. Using a D study, the authors examined the probe length necessary for making low-stakes decisions (defined as reliability $\geq .70$) and high-stake decisions (defined as reliability $\geq .90$). For inter-individual decisions, 1-minute probes were sufficient for low-stakes decisions and probes of 4-minutes were needed for high-stakes decisions. For intra-individual decisions, a 3-minute probe was needed for low-stakes decisions and a 13-minute probe was necessary for high-stakes decisions.

A literature review of articles using PSYCinfo and Google Scholar did not find any generalizability studies on writing CBM. Given the need for further study on the technical adequacy of writing CBM measures, as previously discussed, examining writing CBM using generalizability theory would be a valuable contribution to the literature.

Current Study

Given the advantages of using G theory to assess the technical adequacy of measures through the examination of multiple sources of variance simultaneously, the current study applied G theory techniques to writing CBM. The present study obtained an estimate of variance in writing CBM scores based on persons, occasions, and forms (different story starters), for a composite dependent measure and seven independent scoring methods. The following research questions were addressed:

- How much variance on a composite measure for writing CBM is due to the person (i.e. individual ability), the testing occasion, the specific form being used, and interactions of these facets? Is the composite measure dependable?
- How much variance can be contributed to each scoring method that comprises the composite measure?
- Do different combinations of dependent measures (scoring techniques) provide more dependable outcomes when using writing CBM for both relative and absolute decisions?
- For the composite measure on one occasion, how many probes are necessary to obtain .80 dependability (for low-stakes decisions) and .90 dependability (for high-stakes decisions) for both absolute and relative purposes?
- For each outcome measure considered independently, how much variance on a composite measure for writing CBM is due to the person (i.e. individual ability), the testing occasion, the specific form being used, and interactions of these facets? Is each outcome measure dependable when considered independently of other variables?
- For each outcome measure considered independently of the others, on one testing occasion, how many probe combinations are necessary to obtain .80 dependability (for

low-stakes decisions) and .90 dependability (for high-stakes decisions) for both absolute and relative purposes?

- When considering all of the analyses, is there a benefit to using a composite measure of variables versus individual scoring methods and how do individual scoring methods compare to each other?

It was hypothesized that the largest amount of variance on the composite measure, as well as each measure independently, would be due to persons, with occasions and forms contributing minimal variance. Additionally, it was predicted that the original measures studied at the IRLD (TWW, WSC, CWS), plus correct minus incorrect writing sequences (CIWS) would form a more dependable composite variable compared to the newer measures (total punctuation marks, correct punctuation marks, and words in complete sentences) and that a composite of variables would be more dependable than any one variable considered alone. The original measures (TWW, WSC, and CWS) and CIWS were also predicted to contribute the most to the composite score variance, contributing comparable amounts. The number of forms necessary to obtain reliabilities of .80 and .90 were predicted to vary by measure but average at around three, which is the number of probes often suggested when administering CBM probes.

METHOD

Participants

All consenting students who were present for the study from grades 3 through 5 at a public elementary school in southeast Louisiana were included. There were 91 total participants: 34 in the third grade (37.36% of the total sample), 31 in the fourth grade (34.07%) and 26 in the fifth grade (28.57%). There were 43 boys (47.25%) and 48 girls (52.75%). As done by Hintze, Christ, and Keller (2002), a power analysis was conducted for a repeated measures ANOVA accounting for a large effect size (.40), alpha level of .05, and power of .80. This analysis indicated that only 10 students would be necessary for inclusion in this study to provide suitable power and effect size. Given the availability of the participant pool and the opportunity to both raise power and increase generality, a larger sample was utilized.

Materials

AIMSweb® Writing CBM Probes. AIMSweb® writing CBM probes are a standardized set of probes created out of the research of Deno and Mirkin (mentioned in the introduction). The probes consist of grade-equivalent story starters that can be obtained from the AIMSweb® website. They are scored using total words written (TWW), words spelled correctly (WSC) and correct writing sequences (CWS). The AIMSweb® training manual cites numerous reliability and validity studies indicating that interscorer agreement for all three scoring methods is typically above 90%. Test-retest reliability ranges from .42-.91 for TWW and WSC. Alternate-form reliability measures range from .46-.80 for TWW, WSC, and CWS. Internal consistency as measured by Cronbach's alpha is reported to be .87 for TWW and .70 for WSC (Powell-Smith & Shinn, 2004).

Interscorer Reliability

Interscorer agreement was collected for 19.90% of the collected CBM writing probes. Graduate students previously trained in the scoring of writing CBM were given scoring guidelines to help them independently score the probes using each of the scoring methods. In cases of disagreement, the primary experimenter re-scored the probe and that score was used. To calculate interscorer agreement, each scoring method for the selected probes was compared. The higher score was divided by the lower score and multiplied by 100. For each probe all of the obtained percentages for each scoring method were averaged to obtain a total probe interscorer agreement. All of the total probe interscorer agreement scores were averaged to obtain a total percentage interscorer agreement. The total percentage interscorer agreement equaled 89.99%, ranging from a minimum of 52.4% to a maximum of 100%. Average interscorer agreement for each scoring method is presented in Table 1.

Table 1. Average Interscorer Agreement by Scoring Method.

Scoring Method	Average Interscorer Agreement (Percentage)
Total Words Written	99.29%
Words Spelled Correctly	97.42%
Correct Writing Sequences	92.91%
Correct Minus Incorrect Writing Sequences	81.95%
Total Punctuation	92.07%
Correct Punctuation	87.98%
Words in Complete Sentences	79.51%

Procedural Integrity

Procedural integrity for the administration of writing CBM probes was collected for each administration session through a self-report checklist by the administrator. Total reported integrity averaged 99.65% and ranged from 87.50% to 100%. For 37.50% of sessions an independent observer witnessed the administration session and also recorded integrity. Interrater agreement of integrity averaged 96.76% and ranged from 62.50% to 100%. See Appendix A for the integrity checklist.

Procedure

Consent. Consent was obtained for all students participating in the study. A letter detailing the purposes of the study and study procedures was sent home to the parent of every student eligible to participate in the study. Parents were instructed to sign and return the letter if they gave permission for their child to participate.

Probe administrations. Probes were administered following the procedures detailed in the AIMSweb® training workbook (Powell-Smith & Shinn, 2004). As suggested in the manual, a class wide format was utilized, whereby an administrator gave the probes to an entire class at the same time. Five different graduate students acted as administrators. Each participating classroom was tested individually using the same probes on three separate occasions within a two week time period. The same three probes were given during each testing occasion and were randomly selected from the story starters provided on AIMSweb®. The three probes are presented in Appendices B-D.

Administrators followed the script provided in the AIMSweb® training workbook (Powell-Smith & Shinn, 2004). A copy of the script is presented in Appendix E. For each probe administration, all students were provided a story starter with blank writing space and a pencil. The story starter was read aloud and students were instructed to think about the probe for one

minute. After a minute elapsed, the students were told to begin writing. After 90 seconds, the students were reminded of the content of the story starter. If at any point an individual student paused for longer than 10 seconds or appeared to be finished, he or she was verbally prompted to keep writing. After three minutes, the probes were collected. This process was repeated for each probe.

Scoring. The writing CBM probes were scored seven different ways. The scoring procedures suggested by Deno and Mirkin in the original conception of CBM, and as detailed by the AIMSweb® writing CBM manual (Powell-Smith & Shinn, 2004) were utilized. These procedures are TWW, WSC, and CWS. Four additional measures that have shown promise in recent research on writing CBM (Gansle et al., 2002; Gansle et al., 2004; Gansle et al., 2006; Jewell & Malecki, 2005; McMaster & Campbell, 2008) were used as well. These scoring methods are correct minus incorrect word sequences (CIWS), total punctuation, (TP) correct punctuation (CP), and words in complete sentences (words in CS). These seven measures were chosen based on results of prior research suggesting that they may be useful metrics for scoring writing CBM.

Total words written (TWW). The total number of words written during the 3-minute period was counted. A word was defined as any letter or group of letters that was separated by a space, regardless of spelling or context. Refer to the AIMSweb® manual (Powell-Smith & Shinn, 2004) for specification in regards to hyphens, abbreviations, numbers, and unusual characters.

Words spelled correctly (WSC). In order to score number of words spelled correctly (WSC), the incorrectly spelled words were circled and the number of circles was subtracted from the TWW. Clarification of what is considered correct spelling concerning hyphenation,

capitalizations, abbreviations, contractions, and words with reversed letters can be found in the AIMSweb® manual (Powell-Smith & Shinn, 2004).

Correct writing sequences (CWS). According to AIMSweb® manual (Powell-Smith & Shinn, 2004), the scoring of correct writing sequences (CWS) involves placing a mark ("^") between "two adjacent writing units (words and punctuation) that are correct in the context of what is written" (pg 11). Words and punctuation have to be "mechanically (spelled correctly, appropriate capitalization), semantically, and syntactically correct" (pg 11). The number of marks is then totaled for an estimate of CWS. This procedure was followed for the current study. As acknowledged by the manual, scoring of CWS requires more inference than the other methods so it was recommended to be done carefully. Refer to the manual for numerous examples of CWS.

Correct minus incorrect writing sequences (CIWS). In order to obtain a measure of correct minus incorrect writing sequences, the number of incorrect writing sequences was subtracted from the total number of CWS. To obtain the number of incorrect writing sequences, the number of CWS was subtracted from the total number of possible writing sequences. Negative values were scored as 0.

Total punctuation marks (TP). All punctuation marks used, regardless of whether or not they were used correctly, were counted for this measure. When quotation marks were used, each mark was counted as an individual punctuation mark.

Correct punctuation marks (CP). The number of punctuation marks used correctly was totaled for this measure. If quotation marks were used, each mark was counted as an individual punctuation mark.

Words in complete sentences (W in CS). All words in complete sentences were counted towards this score. A sentence was considered complete if it started with a capital letter, had a subject and a verb, and ended with punctuation.

Data Analysis. Descriptive statistics for the data were calculated for each scoring method by occasions, forms and grades. The mean values were compared to AIMSweb® national norms (2014) for TWW, WSC and CWS (these are the only scoring method for which AIMSweb® provides norms) in order to determine if this sample performed at a level commensurate with national levels. The data was further analyzed for normality through tests of skew and kurtosis performed using the Excel Data Analysis toolpak. Means across occasion, form and grade were compared using one-way Analysis of Variance (ANOVA) tests, with follow-up *t*-Tests for significant ANOVAs. These analyses were performed to test the assumptions that CBM story starters from AIMSweb® comprise exchangeable forms and that they differentially measure performance across grade levels. Additionally, bivariate correlations between the different scoring methods were examined.

A two-facet multivariate G theory analysis was conducted for a composite dependent variable consisting of the seven scoring methods used to score the writing CBM probes. A fully crossed design was used, in which all participants in the analysis were included in all testing occasions and given the same forms. The facets were random as probes and occasions were both randomly sampled from the universe of possible probes and occasions.

The multivariate G theory analysis was conducted using the mGENOVA statistical software program (Brennan, 2001). As suggested by Joe and Woodward (1976), any negative matrix values were set equal to 0. It is not possible to have a negative contribution to variance so any negative contribution was considered to be due to measurement error (Shavelson & Webb,

1991). Matrices of variance and covariance components, generalizability coefficients, and contributions to composite universe score variance from each method were computed. An overall multivariate generalizability coefficient and multivariate phi coefficient were computed for the composite measure. Additionally, generalizability coefficients were computed for three different composite measures consisting of the following combinations of scoring methods: TWW, WSC and CWS; TWW, WSC, CWS and CIWS; TP, CP and W in CS. Following the multivariate G theory analysis, multivariate D studies were conducted to determine the number of forms needed to obtain low and high-stakes relative and absolute decisions for a composite score consisting of all scoring methods given one occasion. One occasion was used since it is most likely that a practitioner will be assessing a student on one occasion and more occasions may not always be available. Combinations of one occasion and various numbers of forms were analyzed (starting with one form and increasing incrementally) until coefficients of at least .80 (for low-stakes decisions) and .90 (for high-stakes decisions) were obtained. Both composite generalizability coefficients (relative coefficients) and composite phi coefficients (absolute coefficients) were obtained for all G and D studies.

Follow-up univariate analyses were conducted for each scoring method using the EduG statistical software package (Cardinet, Johnson & Pini, 2010). Estimates of variance components for each facet, absolute generalizability coefficients, and relative generalizability coefficients were obtained. For each scoring measure, D studies were conducted to find coefficients equal to .80 (for low-stakes decisions) and .90 (for high-stakes decisions) for one occasion. Studies were conducted for the following number of forms: 1, 2, 3, 4, 5, 10, 20, and 50. Both relative and absolute coefficients were calculated for all analyses.

RESULTS

Descriptive Statistics

Descriptive statistics for each scoring method are presented in Table 2. An overall mean and standard deviation are presented, as well as the maximum and minimum value. The average values for TWW, WSC and CWS were within the range of what would be expected based on AIMSweb® norms (2014) for the 50th percentile of third, fourth and fifth grade students. The norms can be referenced in Table 3. Although the average values fell within the expected range, the maximum values far exceeded the provided values for the 90th percentile and the minimum values fell far below the provided values for the 10th percentile. Thus, the range for these scoring methods was larger than might be expected. For CIWS, TP, CP and W in CS no national norms could be found. These values appear to be reasonable for what might be expected based on the norms for the first three scoring methods. These scoring methods also had a large range.

Table 2. Descriptive Statistics by Method.

	TWW	*WSC*	*CWS*	*CIWS*	*TP*	*CP*	*W in CS*
Mean	42.022	38.305	34.634	24.479	4.132	3.874	21.796
Standard Deviation	14.654	14.386	14.772	17.060	3.219	3.024	18.945
Minimum	12.000	8.000	4.000	0.000	0.000	0.000	0.000
Maximum	95.000	94.000	79.000	77.000	24.000	17.000	76.000
Kurtosis	-0.077	-0.075	-0.174	-0.368	2.194	.855	-.826
Skewness	0.448	0.426	0.408	0.493	1.153	.973	.407

To test for normality, kurtosis and skew values were obtained using the Data Analysis toolpak in Excel. These values are also presented in Table 2. Kurtosis and skew values close to zero indicate data approximates a normal distribution (Field, 2009). As suggested by Field (2009), significance tests of skew and kurtosis were not performed since with large samples of 200 or more, these values tend to be significant with very small deviations from normality.

Table 3. AIMSweb® Spring Norms by Grade for Total Words Written (TWW), Words Spelled Correctly (WSC), and Correct Writing Sequences (CWS).

Grade	Percentile	*TWW*	*WSC*	*CWS*
Third	90	59	54	56
	75	49	43	43
	50	39	33	30
	25	30	23	21
	10	23	16	13
	Mean	40	34	32
	Standard Deviation	14	15	16
Fourth	90	66	57	62
	75	56	46	51
	50	45	35	38
	25	35	25	27
	10	25	17	18
	Mean	45	35	39
	Standard Deviation	16	17	17
Fifth	90	74	75	69
	75	63	62	57
	50	51	49	46
	25	41	38	32
	10	31	27	22
	Mean	51	50	45
	Standard Deviation	17	19	18

Values of skewness that are positive indicate more low scores in the distribution and values that are negative indicate more high scores in the distribution. Kurtosis values that are positive indicate a distribution with a high number of scores falling in the middle and wide tails with negative values indicating a flat distribution with light tails (Field, 2009). For all scoring methods skew values were positive. For all scoring methods except for TP and CP kurtosis values were negative. With the exception of TP, the absolute values of all skew and kurtosis values were less than one, suggesting the distributions approximate normal distributions. The absolute values for skew and kurtosis for TP both exceeded one. A visual inspection of the data

graphed as a scatterplot indicated a clear outlier. When this outlier was removed, the skew value dropped to 0.952 and the kurtosis value dropped to 0.774, indicating that the single value was responsible for bringing the skew and kurtosis absolute values above one and causing the data to approximate an abnormal distribution. Although this one data point affected the normality for TP, it was kept in the sample for the univariate and multivariate analyses. This decision was made because none of the other scores for that probe were outlying values, thus it did not make sense to throw out an entire probe (and also an entire participant's data) for one outlying value. It is likely that this outlier reflected imprecision of the scoring method versus abnormal data.

Table 4 further provides the means and standard deviations for each scoring method broken down by forms, occasions and grades. For TWW, WSC, CWS, CIWS and W in CS performance was slightly higher for the second form compared to the other two forms. Performance on the first form for these measures was slightly lower. As shown in Table 4, performance across all measures was marginally higher for the second occasion and was marginally lower for the first occasion when comparing all three occasions. The high standard deviations for all scores indicate that the spread of raw scores for each method was relatively large. Although the scores varied slightly by form and occasion, they fell within one standard deviation of each other.

For each method, an ANOVA was conducted to determine if the differences between the means across forms were significant. Follow-up two-tailed *t*-Tests were conducted when the ANOVA was significant (suggesting that the mean values were significantly different). Given the large number of statistical significance tests and that were performed and thus, the high likelihood of familywise error, a highly conservative alpha value of .001 was used to determine significance. For TWW, the ANOVA was significant, $F(2, 816) = 7.918$, $p < .001$.

Table 4. Mean Scores and Standard Deviations by Method, Form, Occasion and Grade.

	TWW	*WSC*	*CWS*	*CIWS*	*TP*	*CP*	*W in CS*
Form A	39.275[a]	35.692[a]	31.890[a]	21.971	3.982	3.766	20.989
	(14.024)	(13.844)	(14.281)	(16.559)	(3.189)	(3.065)	(18.311)
Form B	44.004	40.059	36.363	25.993	3.751	3.546	22.436
	(14.904)	(14.626)	(15.006)	(17.385)	(3.058)	(2.896)	(19.605)
Form C	42.714	39.165	35.648	25.473	4.663	4.311	21.963
	(14.735)	(14.279)	(14.599)	(16.912)	(3.328)	(3.051)	(18.833)
Occasion 1	38.861[b]	35.582[b]	32.300	22.883	3.886	3.707	21.304
	(14.260)	(13.912)	(14.203)	(16.665)	(2.867)	(2.710)	(18.456)
Occasion 2	44.264	40.234	36.425	25.590	4.319	4.015	22.784
	(14.239)	(13.991)	(14.562)	(17.068)	(3.409)	(3.220)	(19.322)
Occasion 3	42.941	39.099	35.176	24.963	4.190	3.901	21.300
	(14.873)	(14.800)	(15.202)	(17.290)	(3.333)	(3.105)	(18.974)
3rd Grade	33.840[c]	30.428[c]	27.131[c]	18.255[c]	3.389	3.209	17.176
	(10.184)	(10.214)	(10.692)	(12.445)	(2.770)	(2.497)	(14.448)
4th Grade	43.749[c]	39.531[c]	34.660[c]	22.900[c]	3.473	3.241	19.900
	(14.643)	(13.975)	(13.558)	(15.531)	(2.796)	(2.700)	(19.446)
5th Grade	50.718[c]	47.128[c]	44.415[c]	34.500[c]	5.889[c]	5.504[c]	30.098[c]
	(13.916)	(13.904)	(14.995)	(19.318)	(3.540)	(3.376)	(20.730)
Grand Mean and SD	**42.022 (14.654)**	**38.305 (14.386)**	**34.634 (14.772)**	**24.479 (17.060)**	**4.132 (3.219)**	**3.874 (3.024)**	**21.796 (18.945)**

[a] = significantly different from the other two forms ($p < .001$)
[b] = significantly different from the other two occasions ($p < .001$)
[c] = significantly different from the other two grades ($p < .001$)

Follow up *t*-Tests indicated that form A was significantly different from form B, $t(272) = -8.992$, $p < .001$, form A was significantly different from form C, $t(272) = 5.817$, $p < .001$, and forms B and C were not significantly different, $t(272) = -2.667$, $p = .008$. For WSC, the ANOVA was also significant, $F(2, 816) = 7.123$, $p < .001$. Follow up *t*-Tests indicated that form A was significantly different from form B, $t(272) = -7.870$, $p < .001$, form A was significantly different from form C, $t(272) = 5.987$, $p < .001$, and forms B and C were not significantly different, $t(272) = -1.700$, $p = .090$. For CWS, the same pattern was found and the ANOVA was significant, $F(2,$

816) = 7.335, $p < .001$. Follow up *t*-Tests indicated that form A was significantly different from form B, $t(272) = -7.758$, $p < .001$, form A was significantly different from form C, $t(272) = 6.332$, $p < .001$, and forms B and C were not significantly different, $t(272) = -1.261$, $p = .208$. Using the conservative .001 value, the ANOVA for CIWS, TP and CP were not significant, suggesting that mean scores on all forms were not significantly different. The following values were obtained: CIWS: $F(2, 816) = 4.527$, $p = .011$, TP: $F(2, 816) = 7.918$, $p = .011$, and CP: $F(2, 816) = 4.681$, $p = .010$. The ANOVA for W in CS also did not reach significance, $F(2, 816) = .413$, $p = .661$.

As with the forms, for each method, an ANOVA was conducted to determine if the differences between the means across occasions were significant. Follow-up two-tailed *t*-Tests were conducted when the ANOVA was significant (suggesting that the mean values were significantly different). A highly conservative alpha value of .001 was used to determine significance. For TWW, the ANOVA was significant, $F(2, 816) = 10.313$, $p < .001$. Follow up *t*-Tests indicated that occasion one was significantly different from occasion two $t(272) = -9.424$, $p < .001$, occasion one was significantly different from occasion three, $t(272) = 6.264$, $p < .001$, and occasions two and three were not significantly different, $t(272) = -2.445$, $p = .015$. For WSC, the ANOVA was also significant, $F(2, 816) = 7.891$, $p < .001$. Follow up *t*-Tests indicated that occasion one was significantly different from occasion two $t(272) = -8.304$, $p < .001$, occasion one was significantly different from occasion three, $t(272) = 5.666$, $p < .001$, and occasions two and three were not significantly different, $t(272) = -2.080$, $p = .038$. Using the conservative .001 value, the ANOVA for CWS was not significant, $F(2, 816) = 5.661$, $p = .004$. The ANOVAs for CIWS, TP, CP and W in CS were also not significant. The following values

were obtained: CIWS: $F(2, 816) = 1.888$, $p = .152$, TP: $F(2, 816) = 1.299$, $p = .273$, CP: $F(2, 816) = .722$, $p = .486$, and W in CS: $F(2, 816) = .556$, $p = .574$.

A visual inspection of the data in Table 4 by grade shows that the average scores for each method increased from third to fourth grade and again from fourth to fifth grade. These changes in scores show that the average scores on each scoring method increased as the students' grade, and presumably skill level, increased. Such an increase would be expected. As previously mentioned, the average scores for TWW, WSC, and CWS fell close to the 50th percentile provided by AIMSweb® national norms for each grade. To confirm that these differences in scores across grade were significant, an ANOVA was conducted for each method. Follow-up one-tailed *t*-Tests were conducted when the ANOVA was significant (suggesting that the mean values were significantly different). One-tailed tests were used since there was an expected direction that scores would differ (i.e. scores in higher grades would be larger). A highly conservative alpha value of .001 was used to determine significance. The ANOVAs for all of the scoring methods were significant, suggesting that the mean scores across grade level were significantly different from each other. The following values were obtained: TWW: $F(2, 816) = 115.617$, $p < .001$, WSC: $F(2, 816) = 116.621$, $p < .001$, CWS: $F(2, 816) = 116.373$, $p < .001$, CIWS: $F(2, 816) = 72.805$, $p < .001$, TP: $F(2, 816) = 50.784$, $p < .001$, CP: $F(2, 816) = 53.738$, $p < .001$, and W in CS: $F(2, 816) = 35.766$, $p < .001$. Follow up *t*-Tests were performed for each scoring method in order to compare the means between each set of grade levels. For TWW, each of the grade-wise comparisons was significant, suggesting that the mean scores were significantly different for each grade. The following values were obtained: grades 3 and 4: $t(490) = -9.340$, $p < .001$, grades 3 and 5: $t(410) = 15.545$, $p < .001$, and grades 4 and 5: $t(503) = 5.461$, $p < .001$. For WSC, each of the grade-wise comparisons was also significant. The following

values were obtained: grades 3 and 4: $t(490) = -8.920$, $p < .001$, grades 3 and 5: $t(410) = 15.427$, $p < .001$, and grades 4 and 5: $t(503) = 6.126$, $p < .001$. For CWS, each of the grade-wise comparisons was significant, with the following obtained values: grades 3 and 4: $t(490) = -7.397$, $p < .001$, grades 3 and 5: $t(410) = 14.932$, $p < .001$, and grades 4 and 5: $t(503) = 7.650$, $p < .001$. For CIWS, the same significant results for the the grade-wise comparisons were obtained. The following values were obtained: grades 3 and 4: $t(490) = -3.960$, $p < .001$, grades 3 and 5: $t(410) = 11.185$, $p < .001$, and grades 4 and 5: $t(503) = 7.382$, $p < .001$. For TP, the *t*-Test for grades 3 and 4 was not significant, $t(490) = .066$, $p = .474$. The comparison for grades 3 and 5 was significant, $t(410) = 8.286$, $p < .001$, as was the comparison for grades 4 and 5, $t(503) = 8.440$, $p < .001$. The same pattern was found for CP with a non-significant *t*-Test for grades 3 and 4: $t(490) = -.127$, $p = .450$ and significant differences for grades 3 and 5, $t(410) = 8.715$, $p < .001$, and for grades 4 and 5, $t(503) = 8.273$, $p < .001$. *T*-Tests for W in CS had the same pattern as the tests for TP and CP with a non-significant *t*-Test for grades 3 and 4: $t(490) = -1.904$, $p = .029$ and significant differences for grades 3 and 5, $t(410) = 8.126$, $p < .001$ and for grades 4 and 5, $t(503) = 5.697$, $p < .001$

Table 5 depicts bivariate correlations between each scoring method. These correlations were obtained using the Excel Data Analysis toolpak. According to Cohen (1988), an r^2 between .01 and .09 is considered small, between .09 and 0.25 is considered medium, and greater than .25 is considered large. Each scoring method was highly correlated with the other, though TWW, WSC and CWS were more highly correlated with each other than with other methods and the same held for TP, CP and W in CS. These clusters of correlated methods would be expected given what each method measures and the outcomes of prior research. Correct minus incorrect writing sequences (CIWS) appears to fit into both clusters, with correlations greater than .425

with all methods. Given the high correlations, a multivariate generalizability study was conducted prior to conducting independent univariate generalizability studies for each scoring method (as suggested by G. Marcoulides in a personal communication, December 29, 2013).

Table 5. Bivariate Correlations Between Different Scoring Methods.

	TWW	*WSC*	*CWS*	*CIWS*	*TP*	*CP*	*W in CS*
TWW	--	.970	.849	.552	.249	.247	.325
WSC	--	--	.932	.694	.288	.292	.380
CWS	--	--	--	.898	.425	.444	.569
CIWS	--	--	--	--	.438	.469	.635
TP	--	--	--	--	--	.978	.577
CP	--	--	--	--	--	--	.608
W in CS	--	--	--	--	--	--	--

Multivariate Studies

Multivariate Generalizability Studies. Table 6 presents the variance and covariance components obtained in the multivariate G study for each scoring method across each of the facets and interactions. Variance components are presented on the main diagonal and covariance components are presented on the sides. Variance and covariance components can be compared across facets to determine the relative contribution of each facet to the composite score. Across scoring methods, the highest variance and covariance components were due to the persons facet, indicating that the composite score primarily reflects individual ability. The highest covariance components in the persons facet were between TWW, WSC, and CWS, suggesting that these three methods varied closely together. Although not as high, the covariance scores for CIWS and W in CS with the original three methods were also relatively large and all scores across method tended to covary together. The second highest variance and covariance scores were reflective of error.

Table 6. Estimates of Variance and Covariance Components[1] for the Multivariate Generalizability Study.

Source of variation		*TWW (1)*	*WSC (2)*	*CWS (3)*	*CIWS (4)*	*TP (5)*	*CP (6)*	*Words in CS (7)*
	(1)	153.281						
	(2)	147.877	149.865					
	(3)	135.607	146.655	162.120				
Persons (*p*)	*(4)*	103.198	126.799	166.638	207.651			
	(5)	10.729	12.386	18.311	22.723	5.671		
	(6)	9.942	11.768	17.903	22.823	5.375	5.137	
	(7)	70.114	83.812	126.917	164.608	27.966	27.381	202.960
	(1)	7.419						
	(2)	6.230	5.216					
	(3)	5.740	4.809	4.425				
Forms (*f*)	*(4)*	3.736	3.131	2.789	1.687			
	(5)	0.544	0.447	0.430	0.257	0.019		
	(6)	0.352	0.289	0.280	0.153	0.004	0.000	
	(7)	0.958	0.791	0.781	0.250	0.020	0.000	0.000
	(1)	5.665						
	(2)	5.440	5.208					
	(3)	5.587	5.396	5.574				
Occasions	*(4)*	5.109	4.910	5.040	4.484			
(*o*)	*(5)*	0.000	0.002	0.156	0.126	0.187		
	(6)	0.000	0.000	0.081	0.059	0.148	0.114	
	(7)	1.324	1.359	1.316	1.245	0.000	0.000	0.000
	(1)	12.905						
	(2)	11.218	10.132					
	(3)	7.512	7.677	6.680				
Person x	*(4)*	0.994	2.881	4.615	8.691			
Form (*pf)*	*(5)*	0.000	0.012	0.280	0.413	0.434		
	(6)	0.000	0.017	0.346	0.502	0.408	0.389	
	(7)	1.155	2.072	4.383	7.868	3.048	3.339	25.796
	(1)	6.363						
	(2)	5.719	6.078					
Person x	*(3)*	4.755	5.516	5.467				
Occasion	*(4)*	3.749	6.122	7.231	12.168			
(*po*)	*(5)*	0.000	0.000	0.000	0.000	0.574		
	(6)	0.000	0.000	0.000	0.000	0.629	0.676	
	(7)	0.000	0.000	0.000	0.000	0.000	0.000	1.281

Table 6 continued.

Source of variation		*TWW (1)*	*WSC (2)*	*CWS (3)*	*CIWS (4)*	*TP (5)*	*CP (6)*	*Words in CS (7)*
	(1)	0.772						
	(2)	0.612	0.461					
Form x	*(3)*	0.487	0.290	0.249				
Occasion	*(4)*	0.374	0.170	0.105	0.000			
(*fo*)	*(5)*	0.191	0.184	0.099	0.111	0.031		
	(6)	0.239	0.215	0.145	0.157	0.044	0.058	
	(7)	1.672	1.278	1.340	0.949	0.180	0.287	2.551
	(1)	34.369						
	(2)	32.734	36.152					
Person x	*(3)*	30.528	33.442	40.588				
Form x	*(4)*	24.701	30.590	43.187	60.479			
Occasion	*(5)*	1.855	1.566	2.285	1.710	3.562		
(*pfo*)	*(6)*	1.917	1.716	2.387	2.015	3.011	2.868	
	(7)	18.821	18.661	27.783	33.131	5.336	5.338	129.463

[1]Variance components are on the main diagonal, with covariance components on the sides.

Table 7 displays the percentage contribution to universe score variance for the seven different scoring methods. Five of the methods each contributed close to 20% of the universe score variance: CIWS (21.40%), CWS (20.34%), W in CS (18.49%), WSC (17.85%), and TWW (16.57%). The final two scoring methods both contributed less than 3% of the universe score variance with TP contributing 2.71% and CP contributing 2.64%. Table 7 also displays the composite generalizability coefficient (G_{rel}=.95) and the composite phi coefficient (G_{abs}=.94) for the multivariate G study analysis of the composite variable. These values both exceed the value of .90 for high-stakes decisions and .80 for low-stakes decisions, suggesting that using a composite variable of all seven scoring methods (created from three forms across three occasions) provides a dependable method for making both high- and low-stakes relative and absolute decisions.

Table 7. Multivariate Generalizability Study: Percent Contribution from Each Method to Universe Score Variance and Generalizability Coefficients.

Scoring Method	Percentage Contribution to Universe Score Variance
TWW	16.57%
WSC	17.85%
CWS	20.34%
CIWS	21.40%
TP	2.71%
CP	2.64%
W in CS	18.49%
Composite Generalizability Coefficient (G_{rel})	.95
Composite Phi (G_{abs})	.94

Table 8 displays the multivariate G coefficients for three different linear composites of scoring methods. For each combination the methods were weighted equally. For a combination of the three traditional scoring methods, TWW, WSC and CWS, a composite generalizability coefficient (G_{rel}) of .94 and a composite phi coefficient (G_{abs}) of .93 were obtained. The next combination that was analyzed added CIWS to the traditional three methods. For this combination a composite generalizability coefficient (G_{rel}) of .95 and a composite phi coefficient (G_{abs}) of .93 were obtained. The final combination was made from the newer three scoring methods, TP, CP and W in CS. A composite generalizability coefficient (G_{rel}) of .91 and a composite phi coefficient (G_{abs}) of .91 were obtained. These results suggest that composite measures made from any of the three combinations (across three forms and three occasions) would be a dependable estimate of writing ability for both relative and absolute high- and low-stakes decisions.

Table 8. Multivariate Generalizability Coefficients for Different Combinations of Scoring Methods.

$n_o = 3$ $n_f = 3$	Canonical Weights		
TWW	.33	.25	0
WSC	.33	.25	0
CWS	.33	.25	0
CIWS	0	.25	0
TP	0	0	.33
CP	0	0	.33
W in CS	0	0	.33
Composite Generalizability Coefficient (G_{rel})	.94	.95	.91
Composite Phi (G_{abs})	.93	.93	.91

Multivariate Decision Studies. Table 9 displays the results from the multivariate D studies. Both composite generalizability coefficients (G_{rel}) and composite phi coefficients (G_{abs}) are presented for D studies involving one occasion and one, two, three, four and five forms. With one form and one occasion, a composite generalizability coefficient (G_{rel}) of .76 and a composite phi coefficient (G_{abs}) of .73 were obtained. These coefficients did not meet the minimum criteria of .80 or .90 so a D study using one occasion and two forms was conducted. For the D study with one occasion and two forms a composite generalizability coefficient (G_{rel}) of .86 and a composite phi coefficient (G_{abs}) of .83 were obtained. These values exceeded the criterion for low-stakes decisions of .80 but did not meet the criterion for high-stakes decisions of .90. A D study with one occasion and three forms produced a composite generalizability coefficient (G_{rel}) of .88 and a composite phi coefficient (G_{abs}) of .86. Since these values were still below .90, a D study with one occasion and four forms was conducted, resulting in a composite generalizability coefficient (G_{rel}) of .92 and a composite phi coefficient (G_{abs}) of .89. The relative generalizability coefficient was high enough to meet the criterion for high-stakes decisions but the absolute generalizability coefficient was not. A D study with one occasion and

five forms resulted in a composite generalizability coefficient (G_{rel}) of .93 and a composite phi coefficient (G_{abs}) of .90. Since the absolute coefficient met the criterion for high-stakes decisions with this combination of form and occasion, no further D studies were conducted.

Table 9. Multivariate Decision Studies: Generalizability Coefficients.

	$n_o = 1$ $n_f = 1$	$n_o = 1$ $n_f = 2$	$n_o = 1$ $n_f = 3$	$n_o = 1$ $n_f = 4$	$n_o = 1$ $n_f = 5$	$n_o = 3$ $n_f = 3$
Composite Generalizability Coefficient (G_{rel})	.76	.86	.88	.92	.93	.95
Composite Phi (G_{abs})	.73	.83	.86	.89	.90	.94

n_o = number of occasions
n_f = number of forms

Univariate Studies

Univariate Generalizability Studies. Table 10 depicts the percentage of variance components for each facet by method. Additionally, the relative coefficients of generalizability (G_{rel}) and absolute coefficients of generalizability (G_{abs}) for each method are included in the table. For all of the methods, the largest component contributing to variance was the person, indicating that individual ability was largely responsible for the variance among scores. The second highest contributing factor across method was error, which could come from a variety of different sources not directly measured in the study. All of the relative and absolute generalizability coefficients for each method were above .88, suggesting that each of the methods is dependable when using three forms across three occasions. A discussion for the results of each method follows.

Table 10. Univariate Studies: G Coefficients and Percentage of Variance Components by Method.

Facet	TWW	WSC	CWS	CIWS	TP	CP	W in CS
Persons (σ^2_p)	69.4	69.8	72.2	70.3	55.3	55.6	56.0
Occasions (σ^2_o)	3.4	2.4	1.9	0.5	0.1	0.0	0.0
Forms (σ^2_f)	2.6	2.3	2.4	1.6	1.9	1.2	0.0
Person x Occasion ($\sigma^2_{p,o}$)	6.0	4.3	3.0	2.9	4.1	4.2	7.0
Person x Form ($\sigma^2_{p,f}$)	2.9	3.1	2.7	4.1	4.6	7.3	0.4
Occasion x Form ($\sigma^2_{o.f}$)	0.3	0.3	0.1	0.0	0.4	0.6	0.7
Person x Occasion x Form + Residual ($\sigma^2_{p,o,f,e}$)	15.4	17.8	17.7	20.6	33.6	31.0	35.8
Relative Coefficient of Generalizability (G_{rel})	**.94**	**.94**	**.95**	**.94**	**.89**	**.88**	**.90**
Absolute Coefficient of Generalizability (G_{abs})	**.91**	**.92**	**.93**	**.93**	**.88**	**.88**	**.90**

Total words written (TWW). Estimates of variance components for total variance, relative variance, and absolute variance for TWW are presented in Table 11. The persons facet (σ^2_p) contributed the most to the total variance, contributing 69.4%, and the error variance ($\sigma^2_{p, o, f, e}$) was the second largest contributor with 15.4%. The variance attributed to occasions (σ^2_o), forms (σ^2_f), the interaction of person by occasion ($\sigma^2_{p,o}$), person by form ($\sigma^2_{p,f}$), and occasion by form ($\sigma^2_{o,f}$) was much smaller, combining to a total of 15.2%. These results indicate that the bulk of the variance was accounted for by individual variation, followed by error, which was likely due to an unmeasured variable or variables.

As shown in in Table 11, the highest contributor to relative variance was the interaction of person by occasion ($\sigma^2_{p,o}$), which contributed 42.9% and was closely followed by the contribution of error ($\sigma^2_{p, o, f, e}$), which equaled 36.4%. Person by form ($\sigma^2_{p,f}$) contributed a smaller but still significant contribution of 20.8%. These results suggest that when comparing individuals to each other, the bulk of the error variance was accounted for by an interaction of person and the occasion, so possibly certain individuals performed better on certain occasions, which might be expected if attempting to rank individuals across different testing sessions when expecting differential individual growth. Given the closeness of the administration sessions in this study, large amounts of differential growth would not be expected. A large portion of variance was also due to error, which likely reflected variables that were not included in this analysis.

Table 11 also displays the contributions to absolute variance. The highest contributor to absolute variance was also the interaction of person by occasion ($\sigma^2_{p,o}$), which accounted for 30.0%. A large portion of the absolute variance (25.4%) was also due to error ($\sigma^2_{p, o, f, e}$). The contributions of occasions (σ^2_{o}), forms (σ^2_{f}), and the interaction of person and form ($\sigma^2_{p,f}$) were all of comparable amounts, equaling 16.7%, 12.8%, and 14.5% respectively, with the interaction of occasion and form ($\sigma^2_{o,f}$) contributing a negligible amount. These proportions indicate that when comparing a person's performance to their past performance, the interaction of person and occasion accounted for the bulk of the variance, followed by an unmeasured variable (error). The interaction of person and occasion could reflect individual variation on different days or could be indicative of growth, which would be expected of an elementary age student, but not necessarily expected in this study.

Table 11. Total Words Written: Estimates of Variance Components for the Univariate Analysis.

Facet	Degrees of Freedom	Estimated Variance Component	Percentage of Total Variance	Relative Error Variance	Percentage of Relative Variance	Absolute Error Variance	Percentage of Absolute Variance
Persons (σ^2_p)	90	153.270	69.4	--	--	--	--
Occasions (σ^2_o)	2	7.425	3.4	--	--	2.475	16.7
Forms (σ^2_f)	2	5.692	2.6	--	--	1.898	12.8
Person x Occasion ($\sigma^2_{p,o}$)	180	15.302	6.0	4.434	42.9	4.434	30.0
Person x Form ($\sigma^2_{p,f}$)	180	6.445	2.9	2.149	20.8	2.148	14.5
Occasion x Form ($\sigma^2_{o.f}$)	4	0.711	0.3	--	--	0.079	0.5
Person x Occasion x Form + Residual ($\sigma^2_{p,o,f,e}$)	360	33.892	15.4	3.766	36.4	3.766	25.4

As shown in Table 10, the relative coefficient of generalizability (G_{rel}) for TWW was equal to .94 and the absolute coefficient of generalizability (G_{abs}) was equal to .91. These values are both very high, suggesting that TWW was a dependable measure for both high- and low-stakes relative and absolute decisions when measured across three forms over three occasions.

Words spelled correctly (WSC). For WSC, estimates of variance components for total variance, relative variance, and absolute variance are presented in Table 12. The persons facet (σ^2_p) contributed the most to the total variance, contributing 69.8%. The second largest contribution was error ($\sigma^2_{p,\,o,\,f,\,e}$), equaling 17.8%. The variance attributed to the occasions (σ^2_o), forms (σ^2_f), the interaction of person by occasion ($\sigma^2_{p,o}$), person by form ($\sigma^2_{p,f}$), and occasion by form ($\sigma^2_{o,f}$) were much smaller, combining to a total of 12.4%. These numbers suggest that the

bulk of the variance was attributed to individual variation, followed by error, which was likely due to an unmeasured variable or variables.

Table 12 also has estimates for contributions to relative variance. The highest contributor was due to error ($\sigma^2_{p, o, f, e}$), equaling 44.4%. The interaction of person by occasion ($\sigma^2_{p,o}$) closely followed, contributing 32.1%, with person by form ($\sigma^2_{p,f}$) contributing 23.5%. These results suggest that when comparing individuals to each other, the bulk of the variance was due to error, which likely reflected a variable not included in the study.

The contributions to absolute variance are also presented in Table 12. The highest contributor to absolute variance was also error ($\sigma^2_{p, o, f, e}$), contributing 32.7%. This contribution was followed by the interaction of person by occasion ($\sigma^2_{p,o}$), which accounted for 23.6%. The facets of occasions (σ^2_{o}), forms (σ^2_{f}), and the interaction of person and form ($\sigma^2_{p,f}$) all contributed small amounts, equaling 13.3%, 12.7%, and 17.2% respectively, with the interaction of occasion and form ($\sigma^2_{o,f}$) contributing a negligible amount. These proportions indicate that when comparing a person's performance to their past performance, the largest amount of variance was attributed to error, likely in the form of an unmeasured variable or variables. The interaction of person and occasion also accounted for a significant amount of variance, which could have been due to individual variation on different days or could be reflective of growth.

As indicated in Table 10, the relative coefficient of generalizability (G_{rel}) for WSC was equal to .94 and the absolute coefficient of generalizability (G_{abs}) was equal to .92. These values indicate that WSC was a dependable measure for both high- and low-stakes relative and absolute decisions when measured across three forms over three occasions.

Table 12. Words Spelled Correctly: Estimates of Variance Components from the Univariate Analysis.

Facet	Degrees of Freedom	Estimated Variance Component	Percentage of Total Variance	Relative Error Variance	Percentage of Relative Variance	Absolute Error Variance	Percentage of Absolute Variance
Persons (σ^2_p)	90	148.89	69.8	--	--	--	--
Occasions (σ^2_o)	2	5.152	2.4	--	--	1.717	13.3
Forms (σ^2_f)	2	4.916	2.3	--	--	1.639	12.7
Person x Occasion ($\sigma^2_{p,o}$)	180	9.119	4.3	3.040	32.1	3.040	23.6
Person x Form ($\sigma^2_{p,f}$)	180	6.664	3.1	2.221	23.5	2.221	17.2
Occasion x Form ($\sigma^2_{o.f}$)	4	0.547	0.3	--	--	0.061	0.5
Person x Occasion x Form + Residual ($\sigma^2_{p,o,f,e}$)	360	37.874	17.8	4.208	44.4	4.208	32.7

Correct writing sequences (CWS). Table 13 shows estimates of variance components for total variance, relative variance, and absolute variance for CWS. The persons facet (σ^2_p) contributed the most to the total variance, contributing 72.2%. The second largest contribution was due to error ($\sigma^2_{p,\,o,\,f,\,e}$), equaling 17.7%. The variance attributed to the occasions (σ^2_o), forms (σ^2_f), the interaction of person by occasion ($\sigma^2_{p,o}$), person by form ($\sigma^2_{p,f}$), and occasion by form ($\sigma^2_{o,f}$) were each negligible, combining to a total of 10.1%. These numbers suggest that the bulk of the variance was attributed to variation among individuals, followed by error.

Table 13 shows the percent contributions of the different facets to relative variance for CWS. The highest contributor was error ($\sigma^2_{p,\,o,\,f,\,e}$), equaling 50.9%. The interactions of person by occasion ($\sigma^2_{p,o}$) and person by form ($\sigma^2_{p,f}$) contributed comparable amounts of 25.5% and

23.6% respectively. These results suggest that when comparing individuals to each other, the bulk of the variance was due to error, which likely reflects a variable not included in the study. The interactions of person and occasion and person by form contributed a significant amount as well, suggesting that different individuals performed differently across occasions and across forms, affecting their ranking compared to others.

The contributions to absolute variance are also presented in Table 13. The highest contributor to absolute variance was also error ($\sigma^2_{p,\,o,\,f,\,e}$), contributing 37.0%. The contributions of occasions (σ^2_o), forms (σ^2_f), and the interactions of person by occasion ($\sigma^2_{p,o}$), and person by form ($\sigma^2_{p,f}$) were all small but comparable amounts, equaling 11.7%, 15.3%, 18.5%, and 17.2% respectively.

Table 13. Correct Writing Sequences: Estimates of Variance Components from the Univariate Analysis.

Facet	Degrees of Freedom	Estimated Variance Component	Percentage of Total Variance	Relative Error Variance	Percentage of Relative Variance	Absolute Error Variance	Percentage of Absolute Variance
Persons (σ^2_p)	90	160.945	72.2	--	--	--	--
Occasions (σ^2_o)	2	4.146	1.9	--	--	1.382	11.7
Forms (σ^2_f)	2	5.451	2.4	--	--	1.817	15.3
Person x Occasion ($\sigma^2_{p,o}$)	180	6.591	3.0	2.199	25.5	2.197	18.5
Person x Form ($\sigma^2_{p,f}$)	180	6.104	2.7	2.035	23.6	2.035	17.2
Occasion x Form ($\sigma^2_{o.f}$)	4	0.332	0.1	--	--	0.037	0.3
Person x Occasion x Form + Residual ($\sigma^2_{p,o,f,e}$)	360	39.467	17.7	4.385	50.9	4.385	37.0

The interaction of occasion and form ($\sigma^2_{o,f}$) contributed a negligible amount. These proportions indicate that when comparing a person's performance to their past performance, the largest amount of variance was attributed to error, likely in the form of an unmeasured variable or variables. The other variables and interactions all accounted for some of the variance but none more than another so it is difficult to say what factors primarily drove the absolute variation.

As displayed in Table 10, the relative coefficient of generalizability (G_{rel}) for CWS was equal to .95 and the absolute coefficient of generalizability (G_{abs}) was equal to .93. These coefficients indicate that CWS was a dependable measure for both high- and low-stakes relative and absolute decisions when measured across three forms over three occasions.

Correct minus incorrect writing sequences (CIWS). Table 14 depicts estimates of variance components for total variance, relative variance, and absolute variance for CIWS. The persons facet (σ^2_p) contributed the most to the total variance, contributing 70.3%. The second largest contribution was due to error ($\sigma^2_{p,\,o,\,f,\,e}$), equaling 20.6%. The variance attributed to the occasions (σ^2_o), forms (σ^2_f), the interaction of person by occasion ($\sigma^2_{p,o}$), person by form ($\sigma^2_{p,f}$), and occasion by form ($\sigma^2_{o,f}$) were each negligible, combining to a total of 9.1%. These numbers suggest that as with the previous methods, the bulk of the variance was attributed to variation among individuals, followed by an unmeasured variable.

The percent contributions of the different facets to relative variance for CIWS are shown in Table 14. The highest contributor was error ($\sigma^2_{p,\,o,\,f,\,e}$), equaling 49.6%. The interactions of person by occasion ($\sigma^2_{p,o}$) and person by form ($\sigma^2_{p,f}$) contributed significant amounts of 20.8% and 29.6% respectively. These results suggest that when comparing individuals to each other, the bulk of the variance was due to error, which likely represents a variable not included in the

study. The interactions of person and occasion and person by form contributed a significant amount as well.

The contributions to absolute variance for CIWS are also presented in Table 14. The highest contributor to absolute variance was also error ($\sigma^2_{p,\,o,\,f,\,e}$), contributing 43.0%. The interaction of person by form ($\sigma^2_{p,f}$) contributed 25.7% and the interaction of person by occasion ($\sigma^2_{p,o}$) contributed 18.1%. Occasions (σ^2_{o}) and forms (σ^2_{f}) each contributed small amounts, equaling 3.4% and 9.8% respectively. The interaction of occasion and form ($\sigma^2_{o,f}$) did not contribute anything. These proportions indicate that when comparing a person's performance to their past performance using CIWS, the largest amount of variance was attributed to error, likely in the form of an unmeasured variable or variables. Interactions of person and form and person by occasion also contributed to variation in one's performance compared to past performance.

Table 14. Correct Minus Incorrect Writing Sequences: Estimates of Variance Components from the Univariate Analysis.

Facet	Degrees of Freedom	Estimated Variance Component	Percentage of Total Variance	Relative Error Variance	Percentage of Relative Variance	Absolute Error Variance	Percentage of Absolute Variance
Persons (σ^2_{p})	90	207.457	70.3	--	--	--	--
Occasions (σ^2_{o})	2	1.616	0.5	--	--	0.539	3.4
Forms (σ^2_{f})	2	4.627	1.6	--	--	1.542	9.8
Person x Occasion ($\sigma^2_{p,o}$)	180	8.501	2.9	2.834	20.8	2.834	18.1
Person x Form ($\sigma^2_{p,f}$)	180	12.065	4.1	4.022	29.6	4.022	25.7
Occasion x Form ($\sigma^2_{o.f}$)	4	0.00	0.0	--	--	0.00	0.0
Person x Occasion x Form + Residual ($\sigma^2_{p,o,f,e}$)	360	60.664	20.6	6.840	49.6	6.740	43.0

Table 10 shows that the relative coefficient of generalizability (G_{rel}) for CIWS was equal to .94 and the absolute coefficient of generalizability (G_{abs}) equaled .93. Based on these values, CIWS was a dependable measure for both high- and low-stakes relative and absolute decisions when measured across three forms over three occasions.

Total punctuation (TP). Estimates of variance components for total variance, relative variance, and absolute variance for TP are presented in Table 15. The persons facet (σ^2_p) contributed the largest amount to the total variance, equaling 55.3% and the error variance ($\sigma^2_{p, o, f, e}$) was the second largest contributor with 33.6%. The variance attributed to the occasions (σ^2_o), forms (σ^2_f), the interaction of person by occasion ($\sigma^2_{p,o}$), person by form ($\sigma^2_{p,f}$), and occasion by form ($\sigma^2_{o,f}$) were much smaller, combining to a total of 11.1%. These results indicate that the bulk of the variance was accounted for by individual variation, followed by error.

Regarding relative variance, as shown in Table 15, the highest contributor was error ($\sigma^2_{p, o, f, e}$), which equaled 56.3%. The interaction of person by occasion ($\sigma^2_{p,o}$) and person by form ($\sigma^2_{p,f}$) each contributed comparable amounts of 20.8% and 22.9%, respectively. These results suggest that when comparing individuals to each other using TP, the bulk of the variance was accounted for by error, which likely reflects variables that were not included in this analysis.

The contributions to absolute variance are also exhibited in Table 15. Error ($\sigma^2_{p, o, f, e}$) contributed the highest proportion of absolute variance, contributing 50.8%. The interactions of person by occasion ($\sigma^2_{p,o}$) and person by form ($\sigma^2_{p,f}$) contributed moderate and comparable amounts of 18.8% and 20.7% respectively. The contribution of forms (σ^2_f) was 8.7%, with occasions (σ^2_o) and the interaction of occasion and form ($\sigma^2_{o,f}$) each contributing a negligible amount. These proportions indicate that when comparing a person's performance to their past

performance, the bulk of the variance was accounted for by an unmeasured variable or measurement error.

As indicated in Table 10, the relative coefficient of generalizability (G_{rel}) for TP was equal to .89 and the absolute coefficient of generalizability (G_{abs}) was equal to .88. These values suggest that TP was a dependable measure for low-stakes relative and absolute decisions when measured across three forms over three occasions. The generalizability coefficients did not quite meet the criterion for use for high-stakes decisions (equal to .90 or above).

Table 15. Total Punctuation: Estimates of Variance Components from the Univariate Analysis.

Facet	Degrees of Freedom	Estimated Variance Component	Percentage of Total Variance	Relative Error Variance	Percentage of Relative Variance	Absolute Error Variance	Percentage of Absolute Variance
Persons (σ^2_p)	90	5.835	55.3	--	--	--	--
Occasions (σ^2_o)	2	0.009	0.1	--	--	0.003	0.4
Forms (σ^2_f)	2	0.201	1.9	--	--	0.067	8.7
Person x Occasion ($\sigma^2_{p,o}$)	180	0.435	4.1	0.145	20.8	0.145	18.8
Person x Form ($\sigma^2_{p,f}$)	180	0.481	4.6	0.160	22.9	0.160	20.7
Occasion x Form ($\sigma^2_{o.f}$)	4	0.044	0.4	--	--	0.005	0.6
Person x Occasion x Form + Residual ($\sigma^2_{p,o,f,e}$)	360	3.538	33.6	0.393	56.3	0.393	50.8

Correct punctuation (CP). Table 16 presents estimates of variance components for total variance, relative variance, and absolute variance for CP. The persons facet (σ^2_p) contributed the largest amount to the total variance, equaling 55.6% and the error variance ($\sigma^2_{p, o, f, e}$) was the second largest contributor with 31.0%. The variance attributed to occasions (σ^2_o), forms (σ^2_f), the interaction of person by occasion ($\sigma^2_{p,o}$), person by form ($\sigma^2_{p,f}$), and occasion by form ($\sigma^2_{o,f}$) were much smaller, combining to a total of 13.3%. These results indicate that the majority of the variance was accounted for by individual variation, followed by error.

Table 16 also depicts contribution to relative variance for CP, showing that the highest contributor was error ($\sigma^2_{p, o, f, e}$), which equaled 47.2%. The interaction of person by form ($\sigma^2_{p,f}$) contributed the next highest amount with 33.4%, followed by person by occasion ($\sigma^2_{p,o}$) with 19.4%. These results suggest that when comparing individuals to each other, the bulk of the variance was accounted for by error, possibly in the form of an unmeasured variable or reflecting measurement error.

The contributions to absolute variance are also exhibited in Table 16. Error ($\sigma^2_{p, o, f, e}$) contributed the highest proportion of absolute variance, contributing 44.3%. The interaction of person by form ($\sigma^2_{p,f}$) contributed 31.3%, followed by the interaction of person by occasion ($\sigma^2_{p,o}$), which contributed 18.2%. The contribution of forms (σ^2_f) was 5.3%, with occasions (σ^2_o) and the interaction of occasion and form ($\sigma^2_{o,f}$) contributing a negligible amount. These results suggest that when comparing a person's performance to their past performance using CP, the bulk of the variance was accounted for by an unmeasured variable or some other form of error.

Table 16. Correct Punctuation: Estimates of Variance Components from the Univariate Analysis.

Facet	Degrees of Freedom	Estimated Variance Component	Percentage of Total Variance	Relative Error Variance	Percentage of Relative Variance	Absolute Error Variance	Percentage of Absolute Variance
Persons (σ^2_p)	90	5.141	55.6	--	--	--	--
Occasions (σ^2_o)	2	.000	0.0	--	--	.000	0.0
Forms (σ^2_f)	2	0.114	1.2	--	--	0.038	5.3
Person x Occasion ($\sigma^2_{p,o}$)	180	0.393	4.2	0.131	19.4	0.131	18.2
Person x Form ($\sigma^2_{p,f}$)	180	0.675	7.3	0.225	33.4	0.225	31.3
Occasion x Form ($\sigma^2_{o.f}$)	4	0.056	0.6	--	--	0.006	0.9
Person x Occasion x Form + Residual ($\sigma^2_{p,o,f,e}$)	360	2.864	31.0	0.318	47.2	0.318	44.3

As indicated in Table 10, the relative coefficient of generalizability (G_{rel}) and the absolute coefficient of generalizability (G_{abs}) for CP were both equal to .88. These values suggest that CP was a dependable measure for low-stakes relative and absolute decisions when measured across three forms over three occasions. The generalizability coefficients did not meet the criterion for use for high-stakes decisions (equal to .90 or above).

Words in complete sentences (W in CS). Table 17 presents estimates of variance components for total variance, relative variance, and absolute variance for W in CS. The persons facet (σ^2_p) contributed the largest amount to the total variance, equaling 56.0% and the error

variance ($\sigma^2_{p, o, f, e}$) was the second largest contributor with 35.8%. The variance attributed to the interaction of person by occasion ($\sigma^2_{p,o}$) equaled 7.0%, while the variance attributed to occasions (σ^2_o), forms (σ^2_f), the interaction of person by form ($\sigma^2_{p,f}$), and the interaction of occasion by form ($\sigma^2_{o,f}$) were negligible to none These results indicate that the majority of the variance for W in CS was accounted for by individual variation, followed by error.

Contributions to relative variance for W in CS are also presented in Table 17. The highest contributor was error ($\sigma^2_{p, o, f, e}$), which equaled 61.8% and was followed by the interaction between person and occasion ($\sigma^2_{p,o}$) at 36.3%. The interaction of person by form ($\sigma^2_{p,f}$) barely contributed with a value of 1.9%. These results suggest that when comparing individuals to each other, the bulk of the variance was accounted for by error, followed by an interaction between person and occasion, suggesting that individuals performed differently on different occasions for this measure.

The contributions to absolute variance are also exhibited in Table 17. The highest portion of variance was accounted for by error ($\sigma^2_{p, o, f, e}$), which contributed 61.0% and was followed by the interaction between person and occasion ($\sigma^2_{p,o}$), which contributed 35.8%. The contributions of forms (σ^2_f), occasions (σ^2_o), the interaction of occasion and form ($\sigma^2_{o,f}$) and the interaction of person by form ($\sigma^2_{p,f}$) were negligible to none. These results suggest that when making an intra-individual comparison, the bulk of the variance was accounted for by an unmeasured variable or another form of error and by an interaction between person and occasion.

As indicated in Table 10, the relative coefficient of generalizability (G_{rel}) and the absolute coefficient of generalizability (G_{abs}) for W in CS were both equal to .90. These values suggest that W in CS was a dependable measure for both low- and high-stakes relative and absolute decisions when measured across three forms over three occasions.

Table 17. Words in Complete Sentences: Estimates of Variance Components from the Univariate Analysis.

Facet	Degrees of Freedom	Estimated Variance Component	Percentage of Total Variance	Relative Error Variance	Percentage of Relative Variance	Absolute Error Variance	Percentage of Absolute Variance
Persons (σ^2_p)	90	202.807	56.0	--	--	--	--
Occasions (σ^2_o)	2	.000	0.0	--	--	.000	0.0
Forms (σ^2_f)	2	.000	0.0	--	--	.000	0.0
Person x Occasion ($\sigma^2_{p,o}$)	180	25.388	7.0	8.463	36.3	8.463	35.8
Person x Form ($\sigma^2_{p,f}$)	180	1.362	0.4	0.454	1.9	0.454	1.9
Occasion x Form ($\sigma^2_{o.f}$)	4	2.647	0.7	--	--	0.294	1.2
Person x Occasion x Form + Residual ($\sigma^2_{p,o,f,e}$)	360	129.691	35.8	14.410	61.8	14.410	61.0

Univariate Decision Studies. A series of univariate decision studies was conducted for each method to determine how many forms would be needed to obtain relative and absolute coefficients sufficient for low- and high-stakes decisions (.80 and .90 respectively) on one occasion. For each method, the following numbers of forms were tested: 1, 2, 3, 4, 5, 10, 20, and 50. The results are presented in Table 18. For TWW, to make a relative low-stakes decision (i.e. a G_{rel} equal to at least .80) based on one occasion, two forms would be needed and to make a relative high-stakes decision based on one occasion (i.e. a G_{rel} equal to at least .90), 10 forms would be needed. To make an absolute low-stakes decision based on one occasion (i.e. a G_{abs} equal to at least .80), three forms would be needed and even with 50 forms, the G_{abs} never would

reach the target value for high-stakes decisions of .90. For WSC, to make a relative low-stakes decision based on one occasion, two forms would be needed and to make a relative high-stakes decision based on one occasion, 10 forms would be needed. To make an absolute low-stakes decision based on one occasion, three forms would be needed and to make an absolute high-stakes decision based on one occasion, 20 forms would be needed. When using CWS, two forms would be needed to make a relative low-stakes decision based on one occasion and four forms would be needed to make a relative high-stakes decision. To make an absolute low-stakes decision based on one occasion, two forms would be needed and to make an absolute high-stakes decision based on one occasion, 10 forms would be needed. If using CIWS to make a relative low-stakes decision based on one occasion, two forms would be needed and to make a relative high-stakes decision five forms would be needed. To make an absolute low-stakes decision based on one occasion, two forms would be needed and to make an absolute high-stakes decision would require 10 forms. For TP, a relative low-stakes decision based on one occasion would require four forms and to make a relative high-stakes decision based on one occasion, 20 forms would be required. To make an absolute low-stakes decision based on one occasion, five forms would be needed and to make an absolute high-stakes decision based on one occasion, 20 forms would be needed. When using CP for a relative low-stakes decision four forms would be needed and to make a relative high-stakes decision based on one occasion, 20 forms would be needed. To make an absolute low-stakes decision four forms would be required and to make an absolute high-stakes decision 20 forms would be required. If using W in CS to make a relative or absolute low-stakes decision based on one occasion, five forms would be needed. For a relative or absolute high-stakes decision 50 forms would still not be enough to meet the minimum generalizability coefficient requirement of .90.

Table 18. Univariate Decision Study Generalizability Coefficients: One Occasion[1].

1 Occasion	TWW		WSC		CWS		CIWS		TP		CP		Words in CS	
Forms	G_{rel}	G_{abs}	G_{rel}	G_{abs}	G_{rel}	G_{abs}	G_{rel}	G_{abs}	G_{rel}	G_{abs}	G_{rel}	G_{abs}	G_{rel}	G_{abs}
1	.74	.69	.74	.70	.76	.72	.72	.70	.57	.55	.57	.56	.56	.56
2	**.82**	.78	**.83**	.79	**.85**	**.82**	**.82**	**.81**	.70	.69	.70	.70	.69	.69
3	.85	**.81**	.86	**.83**	.88	.85	.86	.85	.77	.76	.77	.76	.75	.74
4	.87	.83	.88	.85	***.90***	.87	.89	.88	**.80**	.79	**.80**	**.80**	.78	.78
5	.88	.84	.89	.86	.91	.88	***.90***	.89	.82	**.82**	.82	.82	**.80**	**.80**
10	***.90***	.86	***.92***	.89	.94	***.91***	.93	***.92***	.87	.87	.87	.87	.84	.84
20	.91	.87	.93	***.90***	.95	.92	.94	.94	***.90***	***.90***	***.90***	***.90***	.86	.86
50	.92	.88	.94	.91	.96	.93	.95	.95	.92	.92	.92	.92	.88	.88

[1]Scores equal to at least .80 are bolded and scores equal to at least .90 are bolded and italicized.

DISCUSSION

The purpose of the current study was to assess the dependability of a composite measure of writing CBM, as well as seven different independent scoring methods, when taking into account the facets of form and occasion. Given the widespread use of CBM as a decision-making assessment tool, it is critical to ensure that an individual's performance on one occasion can be taken as a reliable and valid representation of performance over a larger set of measurements and wider set of contexts. The research on the technical adequacy of different scoring methods for writing CBM has provided mixed and sometimes contradictory results, though the technical adequacy of writing CBM has not been examined through generalizability (G) theory. In a G theory study different sources of error are considered simultaneously, which contrasts classical test theory where each source is considered independently. Thus, the current study was conducted to provide information on the dependability of different measures using G theory techniques. Specifically, seven different research questions were addressed. Each question and a discussion of the respective results and support or non-support for the hypotheses is presented. A broader discussion of results and implications follows.

Discussion of Research Questions

Question 1. How much variance on a composite measure for writing CBM is due to the person (i.e. their individual ability), the testing occasion, the specific form being used, and interactions of these facets? Is the composite measure dependable?

The estimates of variance and covariance components for each facet are presented in Table 6. The sizes of the different components can be compared to each other as an indication of the amount of their relative contribution to overall variance of the composite score. The hypothesis that the largest amount of variance would be due to persons, with forms and occasions contributing a minimal amount of variance, was supported. When comparing estimates of

variance and covariance components across facets, the highest relative contributions are clearly due to the individual, suggesting that the composite score is largely reflective of individual ability. The form that is used, the testing occasion, and the interactions contribute minimal variance. These minimal contributions are ideal considering that CBM should reflect individual ability, which would be expected to be consistent across forms and occasions (although CBM should measure growth, and therefore scores may change by occasion, the occasions in this sample were close enough that it is unlikely that intervention or instruction would show significant effects).

Webb, Shavelson, and Madahian (1983) suggest that the variance and covariance components for persons are one of the most important sources of information that is obtained from a multivariate G study. The variance components for persons represent how much the variance reflects individual ability, while the covariance components provide estimates of covariation between universe scores. Variables with high covariance components vary together, which suggests it is reasonable to form a composite of scores as the measures likely represent the same underlying dimension of the skill being measured. An analysis of the covariance components in this study shows that the relative covariances between TP and CP and the other five measures are smaller compared to the covariances between the other five measures with each other. These small covariances suggest that the relationship between universe scores for TP and CP with the other measures is weak. Given the large covariances between the other measures (TWW, WSC, CWS, CIWS, and W in CS), it seems that it is reasonable to create a composite comprised of them since the scores covary together and likely represent an underlying dimension of writing skills. Given that TP and CP do not covary with the other measures, they might not

represent the same dimension of writing and potentially should not be included in a composite measure along with the other scoring measures.

The second highest variance and covariance components in this study are due to error, which could reflect an unmeasured variable or imprecision in the measure. There are many possible variables that could be driving the high error variance components and future studies should investigate these possible components to determine if the error is random or systematic (i.e. due to a specific variable). One possibility is that the error is reflective of the different ages included in the study. Scores were collapsed across grade and grade was not examined as a facet. It would be reasonable that grade would account for a significant amount of variance given that writing ability should change as students grow, meaning that certain scoring methods may be more appropriate for different age ranges. Also, it is possible that at this particular school the writing instruction for different grades was unique enough that some of the idiosyncrasies of the teacher were reflected as error. For example, maybe one teacher focused significant instruction on the correct use of punctuation so those students paid closer attention to punctuation in their writing compared to other students. It is worth mentioning that although the error variance and covariance components are the second highest contributor, the high composite generalizability coefficient (.95) and the high composite phi coefficient (.94) suggest that the contribution of error is minimal. Since G coefficients are an estimate of the overall variance explained by the contribution of person relative to error, the high coefficients suggest the error is not significant enough to effect interpretation of results.

The third highest variance and covariance components in this study are attributed to the interaction of person and form, which would suggest that individuals are rank-ordered differently across form. Though these components are the third highest, they are still much smaller

compared to the components due to persons. Thus, although there may be some differences in individual rank across form, these differences are likely not very significant. When looking at the covariance components for the interactions of person and form and person and occasion, TWW, WSC, CWS, and CIWS tend to covary together, suggesting that when individuals are ordered differently across form or occasion, they rank similarly across these four methods (i.e. if an individual scores high for TWW, that individual will likely score high for WSC, CWS, and CIWS as well). The low covariance components for TP, CP, and W in CS suggest that individuals are ranked inconsistently for these measures across different occasions and forms (i.e. if an individual scores high on one occasion or form for TP, CP, or W in CS, that individual will not necessarily score high on the other measures).

The composite generalizability component (G_{rel}) is equal to .95 and the composite phi (G_{abs}) is equal to .94. These values suggest that when using three forms across three occasions, the composite measure is highly dependable for both relative and absolute decisions. The percentage of variance attributable to an individual's "true score" is equal to 95% for relative decisions and 94% for absolute decisions.

Question 2. How much variance can be contributed to each scoring method that comprises the composite measure?

Table 7 depicts the percent contribution of each scoring method to the universe score variance for the composite measure. Five methods, TWW, WSC, CWS, CIWS, and W in CS, all contribute approximately 20%. The other two measures, TP and CP, each contribute a negligible amount. The hypothesis for this question was partially supported. It was hypothesized that TWW, WSC, CWS, and CIWS would contribute the most to the composite score variance and that they would contribute comparable amounts. These measures did contribute significantly and contributed comparable amounts, although W in CS also contributed a comparable amount. The

negligible contributions of TP and CP further support that TP and CP might not represent the same dimension of writing as the other measures (as also indicated by the small covariance components). The significant contributions of TWW, WSC, CWS, CIWS, and W in CS further support that these measures likely represent an underlying dimension of writing and that it is reasonable to combine them into a composite measure.

Question 3. Do different combinations of dependent measures (scoring techniques) provide more dependable outcomes when using writing CBM for both relative and absolute decisions?

Table 8 shows the composite generalizability coefficients (G_{rel}) and composite phi coefficients (G_{abs}) for the three different combinations of scoring methods that were analyzed in this study. All of the coefficients for each combination are very high, falling above .91, suggesting that each of the combinations forms a dependable composite when measured across three forms and three occasions. The hypothesis that TWW, WSC, CWS, and CIWS would form the most dependable combination was supported since that combination had the highest coefficients. However, the G coefficient was only slightly higher than the coefficient for the other two combinations. Surprisingly, the composite formed of TP, CP, and W in CS was highly dependable. This outcome is unexpected considering the low contributions of TP and CP to the universe score variance. It is possible that when the other measures are not included, TP and CP contribute more to the total variance and become dependable, but when the other measures are included their contributions diminish. It is also possible that the high generalizability and phi coefficients are primarily due to W in CS and that TP and CP still do not contribute a significant amount. Further analyses would need to be conducted in order to determine what is driving these results.

Question 4. For the composite measure on one occasion, how many probes are necessary to obtain .80 dependability (for low-stakes decisions) and .90 dependability (for high-stakes decisions) for both absolute and relative purposes?

The results of the decision study are depicted in Table 9. On one occasion, two forms are necessary (the average score would be used) in order to obtain a composite generalizability coefficient (G_{rel}) and composite phi coefficient (G_{abs}) of at least .80 for low-stakes decisions. To obtain a G_{rel} sufficient for high-stakes decisions on one occasion, four forms are needed and to obtain a G_{abs} sufficient for high-stakes decisions on one occasion, five forms are needed. The hypothesis for this question was not supported as it was estimated that three forms would be needed. For low-stakes decisions less forms were needed and for high-stakes decisions more forms were needed.

Typically, low-stakes decisions primarily include those for screening, whereby performance is compared among a group of students in order to identify those students who might be at-risk and in need of intervention (often considered the lowest 15-20% in an RTI model). Since these decisions are made by comparing individuals to each other, the G_{rel} is relevant. Using two forms on one occasion for screening purposes is very reasonable, especially considering that writing CBM probes can be administered in a group format. It takes four minutes to administer each probe so even with time to set up, pass out materials, and collect materials, it should take no more than 15 minutes to administer two writing CBM probes to a group of students. This administration could easily be done three times a year as part of a school-wide universal screening program. Hosp et al. (2007) consider progress monitoring of student progress in RTI akin to a screening decision (i.e. a low-stakes decision). Progress monitoring decisions are both relative and absolute, in that a student's individual progress is compared to itself against a personal goal, which is often created based on group norms. For both relative and absolute

decisions, two forms are needed to obtain .80 dependability. Using two forms for progress monitoring is very feasible in a school setting. Although progress monitoring is typically conducted individually, it is possible that a group of students could be administered the writing probes together for progress monitoring purposes, especially considering the story prompts do not depend on skill level (i.e. there is not a different starter for a third grade level writer versus a fourth grade level writer). Thus, using two forms for progress monitoring of writing would be a reasonable expectation within a school setting. It can be assumed that monitoring progress towards an IEP goal is a comparable practice to progress monitoring within an RTI system, thus it can be extended that two forms could also be used for monitoring IEP writing goals.

For high-stakes decisions, which primarily include diagnostic decisions, such as determining if a child qualifies for an exceptionality (such as a learning disability), both relative and absolute comparisons would also be useful. Since the G_{abs} values are lower than the G_{rel} values, and therefore tend to require more forms to reach the .90 dependability criterion, the discussion will focus on the G_{abs} when considering feasibility (thereby acting conservatively by examining the "worst case scenario"). Five forms would be required to meet a .90 dependability with the composite measure for diagnostic purposes. Using five forms on one occasion for such a decision might not be feasible. Although it would be reasonable to take 30 minutes as part of an individual assessment (assuming four minutes for a probe plus time to set up materials), five writing probes would be a high number for an individual to do at one time. The writing probes would also likely be administered as part of a battery of assessments, thus 30 minutes could be a significant time requirement. Curriculum-based measurement (CBM) probes are intended to be administered quickly so taking 30 minutes to administer CBM probes for a single skill contradicts one of the primary features of CBM. Ideally, the measure would be reliable for high-

stakes decisions with less forms and a shorter time requirement. It should be noted that Hosp et al. (2007) recommend the use of a different type of CBM for diagnostic decisions. They recommend the use of a mastery measure, which would focus on a specific set of skills in isolation versus general ability. For example, a mastery measure might solely measure the use of punctuation and would not be intended to represent overall ability. Although the particular writing CBM used in this study might not be suggested for use for diagnostic decisions, it is worthwhile to consider its dependability for diagnostic purposes to help inform possible practice.

It is also worth considering the composite generalizability coefficient (G_{rel}) and composite phi coefficient (G_{abs}) for one form administered on one occasion. It is likely that many practitioners use one form administration to make both high- and low-stakes decisions. Using a composite measure of the methods included in this study based on one form and one occasion yielded a G_{rel} of .76 and a G_{abs} of .73. These coefficients are not high enough to meet the standards for high- and low-stakes decisions used in this study, suggesting that using a composite measure to score a writing CBM administered with one form on one occasion would not be sufficiently dependable (i.e. it would not be acceptable to generalize from the one score to the average score a person would have across all possible testing occasions). Although .80 is usually suggested as a minimum criterion for a coefficient to be considered acceptably dependable (Cardinet, Johnson, & Pini, 2010), other studies use a less stringent criterion for low-stakes decisions, only requiring a .70 coefficient (such as Christ, Johnson-Gros, & Hintze, 2005). Using the lower criterion, a composite measure obtained from one form on one occasion would be considered dependable for low-stakes decisions. Practitioners who wish to use a composite measure for screening purposes when using writing CBM and who are comfortable with using a

minimum criterion of .70 could use only form and one occasion and feel confident that their results are sufficiently dependable.

Question 5. For each outcome measure considered independently, how much variance on a composite measure for writing CBM is due to the person (i.e. individual ability), the testing occasion, the specific form being used, and interactions of these facets? Is each outcome measure dependable when considered independently of other variables?

Table 10 depicts the relative and absolute G coefficients for each scoring method, as well as the percent contributions of each facet to the overall variance. Tables 11-17 depict more detailed components for each method individually. The hypothesis for this question was supported since the persons facet contributed the largest percentage of variance for each method, suggesting that each measure primarily reflects individual ability. Person contributed approximately 70% of the variance for TWW, WSC, CWS, and CIWS. The contribution of person, coupled with the high relative and absolute generalizability coefficients for these methods (all greater than .90), suggest that each of these measures considered independently is highly dependable and largely reflective of individual ability and an individual's "true score".

The second largest contributor for each measure was error. As with the multivariate analysis, it is possible that this large error variance is reflective of differences between grades since scores were collapsed across grades and grade was not included as a facet. Including grade as a facet could possibly account for more of the variance. For TP, CP, and W in CS, the error variable accounted for 30% or more of the variance and persons only accounted for about 55%. These percentages indicate that for these measures, the obtained score is largely reflective of something other than individual ability. This large error component could reflect an unmeasured variable or it could reflect imprecision in the scoring method. Ideally, a measure would primarily reflect individual ability, so these measures should be studied further to determine what else is accounting for the variance. It is also possible that these measures are not optimal measures of

overall writing ability. This possibility seems likely when comparing TP, CP, and W in CS to the other measures, which were all more representative of individual ability as reflected by the higher variance components due to person. Although the persons facet contributes only about half of the variance for TP, CP, and W in CS, the relative and absolute generalizability coefficients for these measures are still above .80, indicating that they are highly dependable when used across three forms and three occasions, particularly for low-stakes decisions.

The facets of forms, occasions, and the interactions contributed minimal variance for all of the scoring methods. This outcome is encouraging, given that the CBM forms are supposed to be equivalent and should provide a comparable measure of skill on parallel occasions (the occasions occurred close enough together that significant growth would not be expected).

Question 6. For each outcome measure considered independently of the others, on one testing occasion, how many probe combinations are necessary to obtain .80 dependability (for low-stakes decisions) and .90 dependability (for high-stakes decisions) for both absolute and relative purposes?

Results from the univariate decision studies for each method are presented in Table 18 and were listed in the results section. Rather than restate each result for each scoring method, this discussion will focus on general trends. The hypothesis for this question was that three forms would be needed to reach the criteria for low- and high-stakes decisions. The hypothesis was supported for some of the scoring measures for relative decisions but it was not supported for any measures for absolute decisions. Low-stakes decisions, such as those involved in screening, typically consist of relative comparisons, thus the G_{rel} can be examined to determine how many forms are needed to obtain a coefficient of at least .80 to inform these decisions given one occasion. For TWW, WSC, CWS, and CIWS, two forms are needed to obtain a G_{rel} of at least .80. This result is the same as the result for the composite score in the multivariate analysis. It is highly feasible to give two writing CBM forms on the same occasion as part of a school-wide

universal screening program given that the probes can be administered in a group and only take four-to-five minutes each. For TP and CP, four forms are needed on one occasion in order to obtain a minimum G_{rel} of .80 and for W in CS five forms are needed. Although administering four or five forms on one occasion in a group format is feasible, it would take a considerable amount of time to score five writing probes for each child in a class, grade, or school. It would also likely be tiring for a student to write that many stories at one time, thus student writing ability might diminish for the later probes. If there are other scoring methods available that have acceptable technical adequacy but can be administered in less time, these methods would be preferable. Absolute generalizability coefficients (G_{abs}) can be examined for intra-individual low-stakes decisions, such as those used for progress monitoring decisions as part of RTI or an IEP. The number of forms needed to obtain a dependability of .80 for the absolute coefficients is generally comparable to those needed for relative decisions. For CWS and CIWS the number stays at two. For TWW and WSC it goes up to three. For CP it is four and for TP and for W in CS it is five. Administration of two or three forms for progress monitoring purposes is feasible, although the more forms needed the less practical it becomes, particularly if a skill is being measured more than one time a week.

The results for the absolute generalizability coefficients (G_{abs}) for high-stakes decisions (i.e. diagnostic decisions) are less encouraging. For TWW and W in CS, even with 50 forms the G_{abs} never reaches the .90 criterion. For WSC, TP, and CP, 20 forms are required on one occasion to reach a G_{abs} of .90 and for CWS and CIWS 10 forms are needed. Administering even 10 forms on one occasion would take the better part of an hour and would likely be tiring for a student. These results potentially suggest that writing CBM scored with one individual method might not

be an optimal measurement tool for diagnostic use, particularly when considering the practicality of administering a high number of forms (10 or more) for both the administrator and the child.

The relative and absolute generalizability coefficients for each method for one form administered on one occasion are important to consider as well. For TWW, WSC, CWS, and CIWS the scores are all close to .70, ranging from .69-.76. For TP, CP, and W in CS the scores range from .55-.57. While none of these scores meet the minimum criterion of .80 for low-stakes decisions, some researchers use a criterion of .70 for these decisions. The scores for TWW, WSC, CWS, and CIWS all meet the .70 criterion. However, the coefficients for TP, CP, and W in CS are significantly below .70, which suggests that using any one of these scoring methods to score one writing CBM probe administered on one occasion would likely not produce a score that is dependable or representative of an individual's "true score" and writing ability.

Question 7. When considering all of the analyses, is there a benefit to using a composite measure of variables versus individual scoring methods and how do individual scoring methods compare to each other?

This question is complicated in that it is affected by many different factors. The hypothesis for this question was that the composite measure would be more dependable than any one measure considered independently. When solely examining dependability, this hypothesis is supported as the results of the generalizability studies yielded a slightly higher generalizability coefficient for the composite measure compared to the individual measures. However, all of the coefficients were very high (between .88 and .95). Differences among the composite measure and the independent variables become more apparent when examining the results of the decision studies. When looking at dependability coefficients for one form administered on one occasion, the coefficients for TWW, WSC, CWS, CIWS, and the composite measure all fall within a similar range (.69-.76), with the coefficients for CWS and the composite falling slightly higher.

However, the coefficients for TP, CP, and W in CS are significantly lower for one form and one occasion, falling between .55 and .57. Additionally, CWS, CIWS, and the composite measure all reach the .80 dependability criterion for low-stakes decisions for both relative and absolute decisions with only two forms administered on one occasion. Total words written (TWW) and WSC also meet this criterion with two forms on one occasion for relative decisions. Total punctuation (TP), CP, and W in CS require at least four forms to meet this criterion. Differences among the scoring methods are even more widespread when looking at the number of forms needed to obtain a .90 dependability as the cut-off for high-stakes decisions. For a relative decision, both the composite and CWS only require four forms on one occasion, with CIWS requiring five. All other measures require at least 10 forms, with W in CS not reaching the .90 criterion even with 50 forms. For absolute decisions, the composite measure reaches .90 dependability with five forms, whereas all of the individual scoring methods require at least 10 forms. Total words written (TWW) and W in CS do not meet the criterion with as many as 50 forms.

A comparison of the utility of the composite measure and each individual scoring method benefits from an examination of other factors in addition to dependability. A brief discussion of such factors for each of the scoring methods and the composite measure follows.

Total words written (TWW). A benefit to using TWW as a scoring method is that it is quick and easy to score. The interscorer agreement for TWW was close to 100%, suggesting that different raters almost always agree on the number of words written in a sample. As indicated in Table 4, the mean scores for TWW were significantly higher for occasion two and occasion three compared to occasion one, as well as for form B and form C compared to form A. These differences could possibly reflect practice effects, as students were able to write more once they

had already been exposed to the task. The forms are intended to be equivalent but the different scores on different forms could indicate that some forms are of differential difficulty as far as word production is concerned. This topic is worth further study since the assumption of form equivalency might be faulty for some story starters.

Words spelled correctly (WSC). Words spelled correctly (WSC) closely mimics TWW in its benefits and drawbacks. Words spelled correctly (WSC) is also relatively simple to score and had a very high IOA in the current study. The significant differences between form A and the other two forms and between occasion one and the other two occasions were also observed.

Correct writing sequences (CWS). Correct writing sequences (CWS) performed the closest to the composite measure in terms of dependability, which suggests that it could be comparable to the composite measure when used independently. However, CWS is one of the most difficult methods to score. Although the interscorer agreement in this study was high, there is a lack of consensus in the literature about how to score CWS. For example, there is no consensus as to whether a comma is scored as a writing unit. The author of the current study created a very detailed scoring procedure for CWS that was based on AIMSweb® criteria and was added to as needed according the sample (for example, the author determined how to score sequences where there was dialogue but no quotation marks). Correct writing sequences (CWS) also requires more inference than other scoring methods. For example, the scorer must determine whether or not a period was omitted between two potentially separate ideas, which affects whether there is an incorrect sequence or not (i.e. was a sentence ended incorrectly or not). Some students will write an entire prompt without using punctuation and will connect ideas with "and". It can be difficult to determine whether or not the student should be penalized for not included separate sentences in a situation where there is no clear ending or beginning to a

sentence. The high interscorer agreement in this study is believed to be a result of the detailed scoring criteria and training procedures that were utilized. Even with very clear criteria, scoring CWS is time-consuming and requires frequent referencing of the scoring rules. This complex scoring procedure clearly contrasts other measures, such as TWW, WSC, or TP, which have very simple and clear scoring criteria that can be easily remembered. The average number of CWS was also significantly less for form A compared to forms B and C. It is possible that CWS scores may partially reflect probe difficulty and not true writing ability.

Correct minus incorrect writing sequences **(CIWS).** In order to obtain CIWS, incorrect sequences are subtracted from correct sequences, thus CIWS has many of the same problems associated with CWS. No detailed standards for this scoring method were found, thus the author had to create standards for this current study. For example, a number written as a number and not spelled out (i.e., "3" instead of "three") does not count as a correct writing unit. It is not clear if a unit inherently becomes an incorrect unit when it is not correct. For example, if a student wrote, "There were 3 alligators." it could either be scored as, "^There^were 3 alligators.^" with 3 CWS (a "^" represents a correct sequence) and 3 CIWS (there correct sequences and no incorrect sequences) or it could be scored as, "^There^were_3_alligators.^" with 3 CWS and 1 CIWS (a "_" represents an incorrect sequence so there would be three correct sequences and two incorrect sequences, thus 1 CIWS). For the purposes of the current study, the latter method was used. Since CIWS involves subtracting incorrect sequences from correct sequences, scoring disagreements as to whether or not a unit is correct become larger disagreements in the CIWS score. It is also counterintuitive to have a measure of writing that could be negative. In this study, negative values were scored as 0 but it is unclear as to whether this practice is common. As with CWS, scoring CIWS requires more inference than other

methods and is more difficult to score. The scoring difficulties inherent in scoring CIWS are reflected in the interscorer agreement for this study, which equaled 81.95%. This score exceeds the minimum criterion for acceptable agreement of 80% (Cooper, Heron, & Heward, 2007), though it is significantly lower than the agreement for most of the other scoring methods. Despite its many disadvantages, CIWS performed well across analyses. It contributed the most to the multivariate composite variance and covaried with all of the other scoring methods. It also increased the generalizability coefficients when added to a composite measure comprised of TWW, WSC, and CWS.

Total punctuation (TP). An advantage of TP is that it is generally easy and quick to score. Although it is relatively simple to score, it was the one scoring method included in this study that had kurtosis and skewness values that suggested abnormality in the data. Upon inspection, there was a significant outlier that seemed to be driving this abnormality. It is not clear if the outlier reflects an abnormal individual ability or reflects imprecision in the scoring method (though that individual did not obtain a score as high on any of the other probes). Total punctuation (TP) also did not significantly differ between grades 3 and 4. If TP does not differ significantly between two grades it might not be an appropriate measure to use for screening purposes to identify students who are at-risk since a student in fourth grade could be writing on a third grade level and this would not be reflected in the score. It also might not be appropriate for progress monitoring growth of writing ability if there is not much difference between the amount of TP produced among different grade levels, suggesting that TP may not be sensitive to change. Additionally, TP contributed minimally to the variance in the multivariate study and did not covary with the other measures. Low covariance scores for interactions of person and occasion and person and form suggest that TP did not vary along with other measures, indicating that if

TP was ranked differently from one occasion to another or across forms, the other measures were not ranked similarly across occasion and form. The contribution of error variance was also high in the univariate analysis, while the contribution attributed to person was low.

Correct punctuation (CP). Correct punctuation (CP) has many of the same attributes as TP, in that it is relatively quick and easy to score. Although it appears simple, CP requires more inference than TP. There might be some situations in which it is not clear if a punctuation mark is correct. For example, the use of commas can be idiosyncratic. In the current study, if a punctuation mark was not immediately recognizable as being incorrect it was counted as correct. As with TP, CP did not differ significantly between grades 3 and 4 and did not contribute substantive variance in the multivariate study. It also did not covary with the other measures and the error variance contributed a large amount relative to the variance due to person in the univariate analysis.

Words in complete sentences (W in CS). As defined in past studies (i.e. Gansle et al., 2002), a word is counted as a W in CS if it is part of a sentence that starts with a capital letter, has a subject and a verb, and ends with punctuation. Although this measure appears to be relatively straight forward, a number of problems arose during the scoring process in the current study. For example, it is unclear how to score the last sentence that a student writes. Often a student ended a writing prompt mid-sentence and thus no words in the last sentence were counted towards this measure. Other students always placed a period at the end of the writing sample when the time was finished. For this measure, students who ended the prompt mid-sentence (and who were following the probe directions by doing so), would be penalized because the words in the last sentence would not be scored. Additionally, some sentences could potentially start with a capital letter, have a subject and a verb, and end with punctuation but not be fully formed. According to

the definition of the scoring criteria, words in this type of sentence would count towards the W in CS score. There were also a number of students who used short exclamatory phrases in their writing, such as "Wow!". Based on the scoring criteria, this phrase would not count as a W in CS. As with CWS, the scorer also had to infer whether or not a run-on phrase should consist of two separate sentences (i.e. did the student fail to place a punctuation mark between two sentences). If a student writes one run-on sentence for the entire sample and the sentence begins with a capital letter, has a subject and a verb, and ends with punctuation, every word written would potentially count as a W in a CS, which does not seem like an accurate representation of the writing sample. The lack of clarity for scoring procedures for W in CS is reflected in the low interscorer agreement, which equaled 79.51%. As mentioned in the introduction, it is important for a measure to have face validity. Although W in CS seems to be a reasonable way to measure writing ability conceptually, the majority of the scorers in the current study judged it to be a flawed measure after using it. Words in complete sentences (W in CS) also did not covary with the other measures for the interactions of person by form and person by occasion and had a high contribution of error variance when considered independently. Additionally, as with TP and CP, W in CS did not significantly differ between grades 3 and 4.

Composite Measure. A composite measure created from various different scoring methods could theoretically be a way to obtain a single measure of writing, thereby allowing for the interpretation of one single score versus the simultaneous interpretation of multiple scores. A potential pitfall to using a composite measure, however, is the ease with which a practitioner could obtain a composite score consisting of a linear combination of variables. In order for a composite measure to be useful, a simple algorithm or a computer scoring program, in which numbers could be plugged in to obtain a composite score, would be necessary. Additionally, any

problems of the individual methods would be compounded for whichever methods are included in the composite.

As previously mentioned, TP and CP have relatively lower covariances with the other measures, suggesting that they might not reflect the same dimension of writing and may not be reasonable to include as part of a composite with the other methods. Although this relationship is somewhat reflected in the lower bivariate correlations, it is important to consider covariance in addition to bivariate correlations. The two factors both support the notion that TP and CP do not appear to act in the same way as the other scoring methods. Total punctuation (TP) and CP also contributed a minimal amount to the variance in the multivariate study, further supporting this notion.

Discussion of Broad Implications

Assessing the dependability of the different writing CBM scoring measures using generalizability theory provides important information as to the degree to which one measurement can generalize to a larger set of measurements. Curriculum-based measurement (CBM) is used almost uniquely for that purpose since an isolated measurement is taken to be a representation of overall functioning for a particular skill. When originally created, CBM was conceptualized as representing "vital signs" of educational ability and success in particular skill areas (Deno & Mirkin, 1977). In order to meet this requirement, CBM needs to be dependable. The technical adequacy of the original measures of writing CBM used in the IRLD studies (TWW, WSC, and CWS) has been questioned in recent studies and newer scoring methods have been proposed. It is important to gain clarity regarding the technical adequacy of all of these measures, especially considering that frequently used progress monitoring programs, such as AIMSweb®, suggest the use of TWW, WSC, and CWS, thus these measures are likely be used

by many practitioners despite conflicting reliability and validity scores in the literature. Generalizability theory techniques have been applied to reading and math CBM but not to writing CBM. The current study was conducted to address all of these issues.

Results of this study suggest that all of the measures, as well as the composite, are highly dependable when using three forms across three occasions for students in grades 3 through 5. When all variables are considered across the multivariate and univariate studies, it appears that the original measures included in the IRLD studies, TWW, WSC, and CWS, still hold the most promise. This outcome contradicts the findings of the more recent studies, which generally resulted in lower reliability and validity coefficients for these measures (Gansle et al, 2004; McMaster & Espin, 2007; Tindal & Parker, 1991). In the current study, each of these measures covaried together, provided a high contribution to the multivariate composite, and had a large portion of variance contributed to the individual when considered alone.

Correct minus incorrect writing sequences (CIWS) also covaried with the original three measures and contributed a high amount to the composite variance. However, CIWS had lower interscorer agreement. With more clarity on scoring criteria, CIWS could potentially be a highly useful measure as it was largely dependable and performed comparably to TWW, WSC, and CWS. CIWS has been suggested as a promising measure of writing in upper elementary students (Wessenburger & Espin, 2005), so clarity as to scoring procedures is needed.

Although W in CS covaried with these measures and contributed to the multivariate composite, there was a low interscorer agreement and persons only contributed 55% of the variance when considered independently. The performance of W in CS was inconsistent across different types of analyses, questioning its utility. Gansle et al. (2004; 2006) also found a lower interscorer agreement for W in CS compared to other scoring methods. However, it was

suggested as a promising measure given its significance in predicting scores on the *Woodcock Johnson—Revised* (WJ—R; Woodcock & Johnson, 1989) Writing Samples subtest using a multiple-regression analysis (Gansle et al., 2004). Although W in CS may be a useful metric, clarification of scoring procedures and further technical adequacy studies are needed before it should be recommended for widespread use.

Total punctuation (TP) and CP did not perform well across analyses. This result contradicts some of the newer writing CBM studies, such as Gansle et al. (2002) and Gansle et al. (2004). However, these studies did not examine the dependability of the measures. Gansle et al. (2002) found high correlations of TP and CP with a teacher rating of writing ability, as well as with the *Iowa Test of Basic Skills* (ITBS). Correct punctuation (CP) was also found to be a significant predicator in a multiple-regression analysis for predicting performance on the ITBS and in a multiple-regression for predicting teacher ranking of writing ability. Additionally, TP had the highest correlation with scores on the *Woodcock Johnson—Revised* (WJ—R; Woodcock & Johnson, 1989) Writing Samples subtest (Gansle et al., 2004) and entered first in a multiple-regression analysis predicting the Writing Samples score. Gansle et al. (2006) suggested that indices of punctuation and sentence production might form a different cluster of writing than TWW, WSC, and CWS. Although this notion was partially supported in this study as TP and CP clustered together, the low dependability scores for TP and CP indicate that these measures should be examined further. If they are not dependable measures, their use may not be supported.

It is important to emphasize that TWW, WSC, CWS, and CIWS hold promise primarily for low-stakes decisions, such as those involved in universal screening and progress monitoring of RTI or IEP goals, specifically with the administration of two or three forms (depending on the scoring method and whether it is a relative or absolute decision) on one occasion. When making

low-stakes decisions, there does not appear to be a benefit to using a composite measure as it would be more time-consuming and would not provide substantially higher dependability compared to the independent use of TWW, WSC, or CWS. Correct writing sequences (CWS) performed the closest to the composite measure across different analyses and slightly outperformed all of the other measures; therefore, it might be a preferable method, although TWW is quicker and easier to score. Considering that CWS has shown to be a promising measure in some of the IRLD studies (McMaster & Espin, 2007; Videen, Deno, & Marston, 1982) and also many of the newer studies (Gansle et al., 2002, McMaster & Campbell, 2008), its use seems to be the most consistently supported and would be suggested by the author of this study.

When considering the use of writing CBM for diagnostic decisions, none of the scoring measures used individually would be acceptable, requiring at least 10 forms to obtain .90 dependability, which is not feasible. A composite measure appears to be better suited for this purpose, although five forms on one occasion would still be required. Three forms across three occasions would also yield a generalizability coefficient of over .90. However, three different testing occasions may not be available for one given student. Given that a diagnostic decision is a high-stakes decision, it is reasonable to require substantive time and effort for administration and scoring. Although, as suggested by Hosp et al. (2007), this type of CBM may not be appropriate to use for diagnostic decisions. The results of this study seem to support that notion as it would require a significant number of forms to use CBM for diagnostic decisions and this might not be feasible, especially considering that this would likely be administered as part of a larger assessment battery. Writing CBM could potentially be used as part of an assessment to provide convergent validity and verify referral concerns. For example, if a child qualifies for a

learning disability in writing based on a standardized achievement test, such as the *Woodcock-Johnson III Test of Achievement, Normative Update* (Woodcock, McGrew & Mather, 2007), the writing CBM scores could be examined to determine if they support this classification. Writing CBM coupled with a standardized achievement test of writing may help provide a more complete picture of an individual's writing ability, as well as provide an indication of the effectiveness of a universal screening program (i.e. are the right children being identified through the screening process).

It is important for practitioners to consider the dependability of the composite and the individual scoring methods when using one form administered on one occasion. It is likely that many practitioners use only one form administered in one session as a measure of progress, which is understandable considering that the progress of multiple students will likely need to be assessed on a weekly or even twice weekly basis and thus quickness might be valued. Some practitioners might be comfortable with the coefficients obtained by the composite, TWW, WSC, CWS, and CIWS for one form administered on one occasion for screening and progress monitoring since they are above .70. However, other practitioners might not be comfortable with the coefficients since they do not reach .80. In that case, two forms should be administered and the average score should be used. The current study used the .80 criterion and thus supports the latter recommendation. Practitioners who wish to use one form on one occasion should proceed with caution as the outcome might not yield an accurate representation of an individual's writing ability.

Limitations

Some limitations of the current study should be considered when interpreting the results. First, the sample was limited to students in grades 3 through 5. It is possible that different measures of writing CBM are differentially appropriate for students of different ages. Results should not be generalized to students in other grades that were not directly included in the present study. Additionally, grade was not specifically measured as a facet and scores were collapsed across grades. It is possible that the large grade range increased the generalizability coefficients that were obtained by increasing the range of scores (McMaster & Espin, 2007). It is also possible that grade contributed to the large variance components due to error but this speculation cannot be determined since grade was not included as a facet. Other sources of variation, such as scorer and method (as a facet), were also not included, thus their contributions are unknown.

Another limitation of the study is that a large number of analyses were conducted. When a large number of analyses are conducted, it becomes more likely that significant results will be obtained (Field, 2009). Additionally, the sample included in the study was much larger than the size suggested by the power analysis. Although the author chose to use a larger sample given the availability of the participant pool and the resulting increase in power, it is possible that using too large of a sample might overestimate variance components due to increased variability (Hintze et al., 2000). Thus, some of the variance components obtained in the study might be inflated.

Threats to the internal validity of the current study include selection bias, the imprecision of some measures, and possible violations of assumptions. In order to be included in the study, parental consent had to be obtained and a student had to be present for all three testing occasions.

It is possible that students who were not proficient in writing or who had anxiety about their writing ability asked their parents to not sign consent to participate in the study, thus individuals with lower writing abilities might not be represented. Although this scenario is possible, a visual inspection of the writing samples, and anecdotal report of school staff do not suggest that this was the case. The imprecision of certain scoring measures, primarily ones that require a significant amount of inference, such as CWS, CIWS, CP, and W in CS, could have threatened the internal validity of the study as well if the measures were not used consistently among raters. The lack of specificity and agreement regarding the various scoring methods for writing CBM is a problem that extends past this particular study and is something that needs to be addressed by the field. There also may be cause for concern that for some of the measures, one of the forms did not appear to provide equivalent scores (lower scores were obtained for Form A for some scoring methods). A basic assumption of writing CBM is that all forms are equivalent and it is possible that this assumption was violated. The assumption of data normality also might have been violated for the TP measure, suggesting that results should be interpreted with caution for that scoring method.

Another limitation is that there are many scoring methods that were not included as a part of this study. A qualitative measure of writing ability was not included. Additionally, none of the production-independent measures, such as percent of WSC and percent of CWS (Jewell & Malecki, 2005), were utilized. The comparisons made in this study between the individual scoring methods and the composite measure cannot be applied to measures that were not included. Thus, when choosing a method of scoring writing CBM, the results of this study cannot inform a decision that involves excluded measures.

Future Directions

While the results of the current study add important information to the research base regarding writing CBM, some problems remain unresolved and should be studied further. The lack of consensus between different studies regarding the technical adequacy of scoring methods still needs to be addressed. This study supported the use of the original measures used in the IRLD studies (TWW, WSC, and CWS) but these measures have not always been supported in more recent studies. As in the IRLD studies, scores were collapsed across grade in the current study, while scores were kept separate for each grade in the newer studies. McMaster and Espin (2007) suggest that this difference in the range of the samples could be responsible for the higher validity measures in the initial studies since including multiple grades would likely increase the range of scores and the sample size and thus increase reliability and validity coefficients. Ideally, a measure used for scoring writing CBM would have sufficient technical adequacy both across grades and within specific grades. Analyzing scores both ways in future technical adequacy studies could show whether collapsing scores across grades is driving the differing reliability and validity coefficients obtained throughout the writing CBM literature. It would be useful to replicate the current study and examine results collapsed across grade compared to results separated by grade.

Researchers and practitioners could also benefit from the creation of clear and consistent scoring definitions for each method. The inconsistency in scoring guidelines can be confusing and may lead to variation in the use of writing CBM. It also may invalidate the comparison of different studies if the measures were not being used in the same way. Thus, the conclusions that can be drawn across studies may be limited. It is essential to create definitive scoring criteria for

different methods as research on writing CBM continues to advance so that results of different studies can be confidently compared and the body of research can become cohesive.

The composite measure should be analyzed through further analyses to determine the best combination of variables. Variables from the current study as well as other variables included in past research should be included. It would also be beneficial to conduct a factor analysis to examine the conclusion that TWW, WSC, CWS, CIWS, and W in CS seem to measure the same dimension of writing and are reasonable to form into a composite. More research on the different clusters of scoring methods is especially needed considering the results suggested by the current study are slightly different than the results obtained by Gansle et al. (2006), particularly regarding W in CS. Gansle et al. (2006) determined that TWW, WSC, and CWS formed a separate cluster from W in CS, whereas W in CS covaried with TWW, WSC, and CWS in the current study.

The composite measure should also be examined for different types of technical adequacy, such as criterion validity and predictive validity. It would be useful to compare the predictive validity of the composite measure to the predictive validity of the independent measures that proved dependable (i.e. TWW, WSC, CWS, and CIWS). It would be particularly useful for practitioners working in schools to know how the different scoring measures can predict writing performance on high-stakes testing and performance regarding Common Core standards.

It would also be useful to investigate the dependability of the composite and various independent measures when considering growth. This examination would be especially pertinent as CBM measures are frequently used to measure growth towards RTI and IEP goals. Future generalizability studies should include different age ranges, to determine whether these results

apply for students outside of grades 3 through 5. Potentially, different scoring methods could be used for different age ranges or different aspects of writing; such as with reading CBM where there are different measures for early literacy skills and different measures for reading fluency versus comprehension (Hosp et al., 2007). It is possible that a similar structure could work for writing CBM. Jewell and Malecki (2005) suggested that TWW, WSC, and CWS become less valid as students enter older grades. There is also support that CIWS is a valid indicator of writing in older students (Jewell & Malecki, 2005; Wessenburger & Espin, 2005). These claims should be investigated further using generalizability analyses.

Another unresolved question involves the high variance components for the error variable of the multivariate and univariate studies (especially for TP, CP, and W in CS). Future generalizability studies should include different facets to help determine what is driving these scores. For example, grade, method, and scorer could be included as facets. It would provide useful information to examine the amount of variance due to scorers who are well-trained in assessment techniques, such as school psychologists, versus scorers with less training in assessment techniques, such as teachers and school staff. Given the widespread use of universal screening and progress monitoring within RTI and special education, it is likely that both types of individuals are scoring CBM probes in schools.

Other measures of writing CBM, such as mastery measures, should also be studied using generalizability theory. The results of this study do not support the use of writing CBM for diagnostic purposes, unless five forms are used on one occasion with a composite measure or three forms across three occasions with the composite or the independent measures. The time required for that many administrations of writing CBM is likely not feasible for practitioners. It would benefit the field to determine a CBM measure of writing that could be used as part of

diagnostic decisions. Potentially, CBM could be used to provide convergent validity for learning disability assessments that use standardized tests of writing achievement or could be used to verify the relationship between screening procedures and learning disability determinations.

CONCLUSIONS

The results of this study are encouraging in that the composite measure and various scoring measures were highly dependable for both inter- and intra-individual decisions. Results suggest that is reasonable to generalize from the use of three writing CBM forms on three measurement occasions to overall writing ability for students in grades 3 through 5. The original measures included in the IRLD studies, TWW, WSC, and CWS, yielded the most consistently positive results across analyses and CIWS holds promise if scoring criteria can be standardized. Additionally, the decision studies showed that the use of these measures, as well as the composite measure, of writing CBM for universal screening, progress monitoring within an RTI system, and monitoring of performance towards IEP goals is supported using one or two forms on one occasion (depending on the personal criteria of the practitioner). These results and suggestions should be taken into account when using writing CBM while our understanding of various scoring methods and uses of writing CBM continues to be studied through empirical research.

REFERENCES

AIMSweb® Writing CBM Probes (2013). Retrieved from www.aimsweb.com.

AIMSweb® National Norms Tables (2014). Retrieved from www.aimsweb.com.

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

Brennan, R.L. (1992). An NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practice*, 27-34.

Brennan, R.L. (2001). *Manual for mGENOVA, Version 2.1* (Iowa Testing Programs Occasional Papers No. 50).

Brennan, R.L. (2010). *Generalizability theory.* New York, NY: Springer.

Brennan, R.L. (2011). Generalizability theory and classical test theory. *Measurement in Education, 24,* 1-21.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying Generalizability Theory using EduG: Quantitative Methodology Series*. New York, NY: Routledge, Taylor & Francis Group.

Christ, T.J., Johnson-Gros, K.N., & Hintze, J.M. (2005). An examination of alternate assessment durations when assessing multiple-skill computational fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychology in the Schools, 42*(6), 615-622.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Coker Jr., D.L. & Ritchey, K.D. (2010). Curriculum-based measurement of writing in kindergarten and first grade: An investigation of production and qualitative scores. *Exceptional Children, 76*(2), 175-193.

Cooper, J.O., Heron, T.E., & Heward, W.L. (2007). *Applied behavior analysis (*2nd ed.). Upper Saddle River, NJ: Pearson.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Deno, S.L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232.

Deno, S.L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37* (3), 184-192.

Deno, S.L. et al. (2009). Developing a school-wide progress-monitoring system. *Psychology in the Schools, 46*(1), 44-55.

Deno, S.L. & Fuchs, L.S. (1987). Developing curriculum-based measurement systems for data-based special education problem solving. *Focus on Exceptional Children, 19*(8), 1-16.

Deno, S.L., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children, 48*(4), 368-371.

Deno, S.L., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. (1982). *The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study* (Vol. IRLD-RR-87). University of Minnesota, Institute for Research on Learning Disabilities.

Deno, S.L., & Mirkin, P.K. (1977). *Data-based program modification: A manual.* Reston, Virginia: Council for Exceptional Children.

Deno, S.L, Mirkin P.K., & 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.

Education for All Handicapped Children Act (1975). PL 94-142.

Field, A. (2009). *Discovering Statistics Using SPSS.* (3rd ed.). London: Sage.

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., Deno, S.L., & Mirkin, P.K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*(2), 449-460.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education, 41*(2), 121-139.

Gansle, K.A., Noell, G.H., VenDerHeyden,, 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*(4), 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*(3), 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*(4), 435-450.

Gresham, F., Reschly, D., & Shinn, M. R. (2010). RTI as a driving force in educational improvement: Research, legal, and practice perspectives. In M.R. Shinn & H.M. Walker (Eds.), *Interventions for achievement and behavior problems in a three-tier model including RTI* (pp. 47-77). Bethesda, MD: National Association of School Psychologists.

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

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

Henderson, M. (2009). *Predicting performance on high-stakes testing: Validity and accuracy of curriculum-based measurement of reading and writing* (Doctoral dissertation). Retrieved from http://www.etd.lsu.edu.

Hintze, J.M. (2005). Psychometrics of Direct Observation. *School Psychology Review, 34*(4), 507-519.

Hintze, J.M., Christ, T.J., & Keller, L.A. (2002). The generalizability of CBM survey-level mathematics assessments: Just how many samples do we need? *School Psychology Review, 31*(4), 514-528.

Hintze, J.M. & Matthews, W.J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral observations. *School Psychology Review, 33*(2), 258-270.

Hintze, J.M., Owen, S.V., Shapiro, E.D., & Daly, E.J. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15*(1), 52-68.

Hintze, J.M. & Pelle Petitte, H.A. (2001). The generalizability of CBM oral reading fluency measures across general and special education. *Journal of Psychoeducational Assessment, 19*, 158-170.

Hosp, M.K., Hosp, J.L., & Howell, K.W. (2007). *The ABCs of CBM: A practical guide to curriculum-based measurement*. New York, NY: The Guilford Press.

H.R. 1350--108th Congress: Individuals with Disabilities Education Improvement Act of 2004. (2003). In GovTrack.us (database of federal legislation). Retrieved October 10, 2012, from http://www.govtrack.us/congress/bills/108/hr1350.

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*(1), 27-44.

Joe, G.W., & Woodward, J.A. (1976). Some developments in multivariate generalizability. *Psychometrika, 41*(2), 205-217.

Lai, C., Park, B.J., Anderson, D., Alonzo, J., & Tindal, G. (2012). *An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 5* (BRT Technical Report No. 1220). Retrieved from University of Oregan, Behavioral Research and Teaching website: http://www.brtprojects.org/publications/technical-reports.

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 children. *Assessment for Effective Intervention 28*(3 & 4), 23-35.

Light, R.J. (2001). *Making the most of college.* Cambridge, MA: Harvard University Press.

Louisiana Department of Education. (2010). *LEAP/GEE technical summary report*. Retrieved from http://www.doe.state.la.us/topics/leap_gee_technical_report.html.

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

Malecki, C.K. & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools, 40*(4), 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 test and direct measurement techniques in measuring pupil progress* (Vol. IRLD-RR-49). University of Minnesota, Institute for Research on Learning Disabilities.

Marston, D., Mirkin, P., & Deno, S. (1984). Curriculum-based measurement: An alternative to traditional screening, referral, and identification. *The Journal of Special Education, 18*(2), 109-117.

McMaster, K. L. & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review, 37*(4), 550-566.

McMaster, K.L., Du, X. & Pétursdóttir, A. (2009). Technical features of curriculum-based measures for beginning writers. *Journal of Learning Disabilities, 42*(1), 41-60.

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

Mercer, S.H., Dufrene, B.A., Zoder-Martell, K., Lestremau Harpole, L., Mitchell, R.R., & Blaze, J.T. (2012). Generalizability theory analysis of CBM maze reliability in third- through fifth-grade students. *Assessment for Effective Intervention, 37*(3), 183-190.

National Center on Student Progress Monitoring, U.S. Office of Special Education Programs. Retrieved September 17, 2012 from www.studentprogress.org

National Commission on Writing for America's Families, Schools, and Colleges. (2003, April). *The Neglected "R": The need for a writing revolution*. Retrieved from http://www.host-collegeboard.com/advocacy/writing/

National Commission on Writing for America's Families, Schools, and Colleges. (2008, April). *Writing, technology and teens*. Retrieved from http://www.host-collegeboard.com/advocacy/writing/

No Child Left Behind Act of 2001, 20 U.S.C. 70 § 6301 *et seq.* (2002).

Olinghouse, N.G. & Santangelo, T. (2010). Assessing the writing of struggling learners. *Focus on Exceptional Children, 43*(4), 1-27.

Parker, D.C., McMaster, K.L., Medhanie, A., & Silberglitt, B. (2011). Modeling early writing growth with curriculum-based measures. *School Psychology Quarterly, 26*(4), 290-304.

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

Powell-Smith, K.A., & Shinn, M.R. (2004). *Administration and scoring of written expression curriculum-based measurement (WE-CBM) for use in general outcome measurement*. Pearson.

Quenemoen, R., Thurlow, M., Moen, R., Thompson, S., & Morse, A. B. (2004). *Progress monitoring in an inclusive standards-based assessment and accountability system* (Synthesis Report 53). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Shapiro, E.S., Hilt-Panahon, A., & Gischlar, K. L. (2010). Implementing proven research in school-based practices: Progress monitoring within a response-to-intervention model. In M.R. Shinn & H.M. Walker (Eds.), *Interventions for achievement and behavior problems in a three-tier model including RTI* (pp. 175-192). Bethesda, MD: National Association of School Psychologists.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: SAGE Publications.

Shinn, M.R. (2010). Building a scientifically based data system for progress monitoring and universal screening across three tiers, including RTI using curriculum-based measurement. In M.R. Shinn & H.M. Walker (Eds.), *Interventions for achievement and behavior problems in a three-tier model including RTI* (pp. 259-292). Bethesda, MD: National Association of School Psychologists.

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.

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. (1991). Identifying measures for evaluating written expression. *Learning Disabilities Research and Practice, 6*(4), 211-218.

U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 and 2007 Writing Assessments, retrieved April 15, 2012, from the NAEP Data Explorer (http://nces.ed.gov/nationsreportcard/nde/). (This table was prepared May 2008.)

Videen, J., Deno, S., & 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.

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

Wayman, M.M., Wallace, T., Wiley, H.I., Tichá, R., Espin, & Espin, C.A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2). 85-120.

Webb, N.M., & Shavelson, R.J. (1981). Multivariate generalizability of general education development ratings. *Journal of Educational Measurement, 18*(1), 13-22.

Webb, N.M., Shavelson, R.J., & Haertel, E.H. (2006). Reliability coefficients end generalizability theory. In C.R. Rao & S. Sinharay (Eds.), *Handbook of Statistics, Vol. 26: Psychometrics* (pp. 81-124). Oxford, U.K.:Elsevier B.V.

Webb, N.M., Shavelson, R.J., & Maddahian, E. (1983). Multivariate generalizability theory. *New Directions for Testing and Measurement, 18*, 67-81.

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 psychoeducational battery* (Rev. ed). Allen, TX: DLM.

Woodcock, R.W., McGrew, K.S., & Mather, N. (2007). *Woodcock-Johnson III, Tests of Achievement Normative Update*. Riverside Publishing.

APPENDIX A: TREATMENT INTEGRITY CHECKLIST

Date:

Class:

For each item, circle either "yes" or "no".

1. The administrator provided all students with a pencil and the appropriate story starter on a sheet of lined paper.	YES	NO
2. The administrator read the directions as written on the AIMSweb® script.	YES	NO
3. The administrator gave the students 1 minute to think about the story starter.	YES	NO
4. After 30 seconds, the administrator repeated the prompt.	YES	NO
5. The administrator told the class to begin writing as instructed on the AIMSweb® script.	YES	NO
6. The administrator walked around the room, prompting students who had stopped writing for 10 seconds to continue writing.	YES	NO
7. After 90 seconds, the administrator repeated the story starter.	YES	NO
8. After 3 minutes, the administrator instructed the students to stop writing and collected the writing prompts.	YES	NO

APPENDIX B: AIMSWEB® WRITING PROBE, FORM A

A person of super-human strength landed in the middle of town and…

APPENDIX C: AIMSWEB® WRITING PROBE, FORM B

We were paddling on a beautiful lake in the woods when our canoe tipped over and…

APPENDIX D: AIMSWEB® WRITING PROBE, FORM C

Every day after school my friends and I would go to the playground and…

APPENDIX E: WRITING CBM SCRIPT (FROM AIMSWEB®)

Written Expression Curriculum-Based Measurement (WE-CBM) Standardized Directions

1. Select an appropriate story starter.

2. Provide the student with a pencil and a sheet of lined paper.

3. Say these specific directions to the student:

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? (Pause). Put your pencils down and listen.

For the next minute think about… "(insert story starter)"

4. After reading the story starter, begin your stopwatch and allow 1 minute for students to "think." (Monitor students so that they do not begin writing).

After 30 seconds say: ***You should be thinking about*** (insert story starter).

5. At the end of 1 minute say: ***Now begin writing.*** Restart your stopwatch.

6. Monitor students' participation. If individual students pause for about 10 seconds or say they are done before the test is finished, move close to them and say ***Keep writing the best story you can.*** This prompt can be repeated to students should they pause again.

7. After 90 seconds say: ***You should be thinking about*** (insert story starter).

8. At the end of 3 minutes say: ***Stop. Put your pencils down.***

If students want to finish their story, it is allowable to do so as long as they complete it on a separate piece of paper.

APPENDIX F: INSTITUTIONAL REVIEW BOARD APPROVAL

Institutional Review Board
Dr. Robert Mathews, Chair
131 David Boyd Hall
Baton Rouge, LA 70803
P: 225.578.8692
F: 225.578.6792
irb@lsu.edu | lsu.edu/irb

ACTION ON PROTOCOL APPROVAL REQUEST

TO: Frank Gresham
Psychology

FROM: Robert C. Mathews
Chair, Institutional Review Board

DATE: November 27, 2012
RE: **IRB#** 3333

TITLE: Multivariate Generalizability of Writing Curriculum-Based Measurement (CBM): An Examination of Form, Occasion, and Scoring Method

New Protocol/Modification/Continuation: New Protocol

Review type: Full ___ Expedited X **Review date:** 11/28/2012

Risk Factor: Minimal X Uncertain _____ Greater Than Minimal_____

Approved X **Disapproved**_____

Approval Date: 11/28/2012 **Approval Expiration Date:** 11/27/2013

Re-review frequency: (annual unless otherwise stated)

Number of subjects approved: 200

Protocol Matches Scope of Work in Grant proposal: (if applicable)_____

By: Robert C. Mathews, Chairman

PRINCIPAL INVESTIGATOR: PLEASE READ THE FOLLOWING –
Continuing approval is CONDITIONAL on:

1. Adherence to the approved protocol, familiarity with, and adherence to the ethical standards of the Belmont Report, and LSU's Assurance of Compliance with DHHS regulations for the protection of human subjects*
2. Prior approval of a change in protocol, including revision of the consent documents or an increase in the number of subjects over that approved.
3. Obtaining renewed approval (or submittal of a termination report), prior to the approval expiration date, upon request by the IRB office (irrespective of when the project actually begins); notification of project termination.
4. Retention of documentation of informed consent and study records for at least 3 years after the study ends.
5. Continuing attention to the physical and psychological well-being and informed consent of the individual participants, including notification of new information that might affect consent.
6. A prompt report to the IRB of any adverse event affecting a participant potentially arising from the study.
7. Notification of the IRB of a serious compliance failure.
8. SPECIAL NOTE:

**All investigators and support staff have access to copies of the Belmont Report, LSU's Assurance with DHHS, DHHS (45 CFR 46) and FDA regulations governing use of human subjects, and other relevant documents in print in this office or on our World Wide Web site at http://www.lsu.edu/irb*

VITA

Katherine Chenier is a native of Louisiana, hailing from the city of New Orleans. She completed her undergraduate education at Tufts University in the cold, northern city of Medford, Massachusetts. She graduated from Tufts with a Bachelors of Arts in Psychology in May 2008. During her time at Tufts, she was able to study abroad in Florence, Italy for a semester to focus on her minor of art history. After her worldly undergraduate experience in Europe and the northern United States, she returned to the warm and humid state of Louisiana. There, she taught special education at an elementary school in New Orleans for a year, during which she decided she would benefit from spending more time on the learning end of the student-teacher relationship. She chose to further her education at Louisiana State University in Baton Rouge, Louisiana, where she has pursued a doctorate in the discipline of school psychology. Katherine currently is working as a school psychologist in a charter school in New Orleans, where she resides with her husband, Jeff, and her dog, Nacheaux.