ABSTRACT

Title of dissertation: MEASURES OF WRITING SKILLS AS PREDICTORS OF HIGH STAKES ASSESSMENTS FOR SECONDARY STUDENTS

Karen Anne Jones, Doctor of Philosophy, 2007

Dissertation directed by: Professor Sylvia Rosenfield
Department of Counseling and Personnel Services

This study examined the potential utility of written expression scoring measures, developed in the curriculum-based measurement research, to monitor student progress and predict performance on a high stakes state mandated assessment for high school students. In response to a teacher generated prompt, 10th-grade students completed 3 brief constructed response (BCR) and 2 extended constructed response (ECR) writing samples throughout the academic year. Writing samples were scored for total words written (TWW), words spelled correctly (WSC), correct writing sequences (CWS), correct minus incorrect writing sequences (CMIWS), percentage of words spelled correctly (%WSC), percentage of correct writing sequences (%CWS), production dependent index, and production independent index. The average time to score a BCR for TWW, WSC, CWS, and CMIWS was over 7 minutes, and the average time to score an ECR was over 16 minutes. Alternate form reliability correlation coefficients between

scoring measures were only in the weak to moderate range. Results revealed that girls wrote more words, spelled more words correctly, produced more correct writing sequences, and produced more correct minus incorrect writing sequences. Across writing samples, statistically significant but small increases were found on scoring measures. Results of multiple regression and logistic regression analyses failed to provide a model that accurately predicted student outcomes.

MEASURES OF WRITING SKILLS AS PREDICTORS OF HIGH STAKES ASSESSMENTS FOR SECONDARY STUDENTS

by

Karen Anne Jones

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
Of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:

Professor Sylvia Rosenfield, Chair
Professor Chan Dayton
Professor Deborah Speece
Associate Professor William Strein
Special Member Deborah Nelson

Dedication

The writing of a dissertation can be a lonely and isolating experience. I could not have accomplished this without the personal and practical support of numerous people who have shared the best and worst moments of this journey with me.

I dedicate this dissertation to my family and friends, especially….

to Darrell who shared in the many uncertainties, challenges, and sacrifices

to Dad, Mom, and David for their endless patience and understanding

to Sonja, Ricia, Mary, Beth, Lindsey, Gina, and Jasmin for their professional and personal support

to Larry who started me on this path and believed in my potential

to Dr. Sylvia Rosenfield for her expert guidance and advice

to others who go unnamed and held my hand at some point along the way.

Table of Contents

List of Tables

Measures of Writing Skills as Predictors of High Stakes Assessments with Secondary Students

Passage of the No Child Left Behind Act (NCLB) of 2001 has led to extensive changes in the education of students in pre-kindergarten through high school. NCLB introduced new requirements intended to raise the achievement of all students through increased accountability and an emphasis on doing what works based on scientific research. Accountability is measured through the requirement that every state implement annual assessments in reading and math in grades 3 through 8 and at least once in grades 10 through 12. The data from the assessments are used to determine if schools and school districts have achieved adequate yearly progress goals.

While NCLB does not require states to attach "high stakes" consequences for students to the assessment results, many states have increased their graduation requirements to include passing the assessments. Beginning with the graduating class of 2009, high school students in the state of Maryland are required to pass four content area assessments, English 2, Algebra, Biology, and American Government to graduate with a high school diploma. Due to the potential negative long-term consequences to students of not passing the assessments, research-based methods are needed to guide teacher instruction and measure student progress toward passing the assessments (Weissenburger & Espin, 2005).

Educational policy continues to emphasize the importance of literacy for all students (No Child Left Behind Act, 2001). Literacy has two components, reading and writing, but many students find writing to be a difficult and frustrating task (McMaster & Espin, 2007). Data from the National Assessment of Educational Progress (U.S.

Department of Education, 2003) provides evidence that many students experience difficulty mastering writing tasks. Three out of every four 4th-, 8th-, and 12th-grade students demonstrated only partial mastery of necessary writing skills and knowledge at their respective grade levels and only 1 in 100 students demonstrated "advanced" writing skills. Providing progress monitoring in the area of writing is therefore an important goal, and one that this study is designed to address.

Progress Monitoring Methods

Teachers assess student performance to achieve a variety of goals (Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006). Gansle et al. (2006) argued that in this time of increased accountability, "one of the most critical functions of educational assessment is to monitor student progress and make instructional decisions" (p. 436). While traditional assessment approaches (i.e. standardized commercial achievement tests) are psychometrically sound, they do not provide useful data required for progress monitoring and instructional decision-making (Shinn, Rosenfield, & Knutson, 1989). In 1974, Carver argued that there are two types of tests: psychometric tests and edumetric tests. Psychometric tests focus on measuring between-individual differences while edumetric tests focus on measuring within-individual growth. He stated that teacher made tests usually focus more on the edumetric dimension than on the psychometric dimension. While tests are usually evaluated according to psychometric principles, they can also be evaluated on edumetric principles. The limitations of traditional assessments led to the development of assessments that integrate school curriculum and instructional goals (Gansle et al., 2006; Shinn et al., 1989).

Academic assessments for the purpose of measuring student progress and informing instruction have a long history in education. Curriculum-based assessment generally refers to "any approach that uses direct observation and recording of a student's performance in the local school curriculum as a basis for gathering information to make instructional decisions" (Deno, 1987, p.41). Four major models of CBA exist in the literature: Curriculum-Based Assessment for Instructional Design (CBA-ID), Criterion-Referenced Curriculum Based Assessment (CR-CBA), Curriculum-Based Evaluation (CBE), and Curriculum-Based Measurement (CBM). A brief summary of the models follows.

Curriculum-based assessment for instructional design is defined as a "methodology used to determine the instructional needs of students based on their performance within the existing course content" (Gickling & Thompson, 1985, p. 217). The major focus of CBA-ID is on instructional planning and the goal is to ensure that the student is placed appropriately within the instructional materials being used (Gickling, Shane, & Croskery, 1989; Gickling & Thompson). The instructional task must have an appropriate amount of challenge for the student, while ensuring that the student possesses the basic skills to be successful. If there is an appropriate match between the student and the instructional task, then the amount of "academic learning time" and time spent on task will be increased (Gickling & Thompson).

Criterion-referenced curriculum based assessment is defined as "the practice of obtaining direct and frequent measures of a student's performance on a series of sequentially arranged objectives from the curriculum used in the classroom" (Blankenship & Lilly, 1981, p.81). The critical feature is to link the assessment to the

local curriculum and instruction (Blankenship, 1985). This assessment includes many areas including basic skills, content areas, and general learning. Due to the incorporation of domains beyond basic skills, the assessment can provide formative and summative data. In order for the assessment to be comprehensive, it would contain items from the beginning, middle, and end of the material taught.

Curriculum-based evaluation is based on the tenant "test what you teach and teach what you test" (Howell & Morehead, 1987, p. 74). The primary goal is to provide information about what skills to teach the student and provide information about the content of instruction. Assessments are constructed to test the subskills that a student needs to be successful in the curriculum. Then, errors made by the student are analyzed and an intervention plan is developed to teach the missing subskills.

Curriculum-based measurement was originally developed as a simple set of standardized procedures teachers can use to monitor student growth and document progress over time in many areas of academic skills (Deno, 1985, 1987). Over two decades of research, primarily at the elementary level, has identified reliable and valid indicators of academic performance in the areas of reading, spelling, written expression, and arithmetic (e.g., Deno, 1985; Deno, Marston, & Mirkin, 1982; Foegen & Deno, 2001; Marston, 1989; Shinn, Ysseldyke, Deno, & Tindal, 1986). Curriculum-based measurement relies on the use of direct, frequent, and efficient measurement of student growth using performance indicators. If the measures are to be administered on a frequent basis for progress monitoring, they must be time efficient, easy to administer and score, easy to understand, and valid for the purpose intended. A substantial body of research supports CBM as a method to gather student performance data to support a wide range of

educational decisions (Deno, 2003). Curriculum-based measurement research has provided data showing its effectiveness in many areas including: improving instruction through the use of goal setting, progress monitoring, and evaluating the effects of change (Fuchs, Deno, & Mirkin, 1984; Fuchs, Fuchs & Hamlett, 1989), developing local performance norms (Marston & Magnusson, 1988), evaluating pre-referral interventions (Shinn, 1995), offering alternative special education identification procedures (Marston & Magnusson; Marston, Mirkin, & Deno, 1984), screening and placement of students (Fewster & MacMillan, 2002), assessing content area learning for secondary students (Ketterlin-Geller, McCoy, Twyman, & Tindal, 2006; Tindal & Nolet, 1995), and diagnostic analysis to adapt instruction (Fuchs, Fuchs, Hosp, & Hamlett, 2003).

Changes in educational policy and increased attention to accountability and high stakes assessments have only served to highlight the critical need for research-based methods to monitor student progress and inform instruction. Of the four CBA models presented, the overwhelming majority of research has been conducted on CBM. While CBM does not provide detailed information to inform instruction, it does provide simple, reliable measures to identify students who may need extra assistance and to monitor progress (Gansle et al., 2006).

Research on the various curriculum-based assessment methods has been largely conducted by CBM researchers. Compared to the amount of research on CBM in reading and math, relatively little research attention has been paid to CBM in writing (Fewster & MacMillan, 2002; Fuchs & Fuchs, 1997; Marston, 1989; Tindal & Parker, 1989). However, recently there has been increased focus on monitoring student performance and progress in writing (Gansle et al., 2006; McMaster & Espin, 2007).

Measures for Evaluating Written Expression

The comparatively small amount of research conducted on written expression may be due to the "special challenges" that the assessment of writing skills presents to researchers (Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002). These challenges include the complexities involved in the direct assessment of writing (Tindal & Parker, 1989), the wide range of possible responses to writing tasks, and the wide range of possible scoring rubrics (Gansle et al., 2002). Due to the infinite number of correct written responses that could be produced by students, objectively scoring the differences in quality can be overwhelming (Gansle et al., 2006).

In 2004, Espin, Weissenburger, and Benson provided a review of different measures for directly evaluating written expression. Each measure reviewed presents advantages and disadvantages and differs in purpose (Espin et al., 2004; Tindal & Parker, 1991). Holistic scoring emphasizes the general impression that the writing gives to the rater and can be used to screen students based on overall writing ability (Tindal & Parker, 1991). Primary trait scoring is criterion based and different scoring guidelines are developed for different writing purposes (Espin et al., 2004). Analytic scoring emphasizes the quality of the writing on several characteristics and the same scoring criteria are used for all writing purposes. Based on their review, Espin et al. (2004) concluded that holistic, primary trait, and analytic scoring measures are not "designed to provide reliable and valid progress information useful for systematic evaluation of instruction" (p. 59). In contrast, curriculum-based measurement was designed to measure progress and to be sensitive to small changes in performance (Deno, 1985).

The purpose of this study was to investigate the use of authentic writing material collected by a classroom teacher to monitor student progress in writing and predict student performance. The writing samples collected were based on the curriculum and written in the format required for an end-of-year high stakes assessment. While the writing samples analyzed in this study were not collected using traditional CBM timed probes, the purpose of the writing was to monitor student progress over time. Further, CBM scoring measures and indices for the assessment of written expression were used to score the writing samples to determine if they would be appropriate for monitoring progress over time in teacher collected writing samples. Since the literature base on CBM in writing research is the most closely related with the purpose of this study, a discussion of this research base will be presented here.

Research on CBM in writing has demonstrated the reliability and validity of different measures to objectively evaluate student writing (Deno et al., 1982; Espin et al., 2000; Gansle et al., 2002; Tindal & Parker, 1989). According to Gansle et al. (2006), while CBA and CBE models have attempted to produce other writing measures to objectively evaluate student writing, the number of studies conducted and the reliability and validity of these measures has been very limited.

Written Expression Scoring Measures

Almost three decades of research has identified reliable and valid scoring measures to monitor student progress in written expression (Deno, Marston, & Mirkin, 1982; Espin et al., 2000; Tindal & Parker, 1989). Research has validated three categories of scoring measures: production dependent indices or fluency measures, production

independent indices or accuracy measures, and an accurate production index (Jewell & Malecki, 2005).

Production dependent measures are referred to as measures of writing fluency because they depend upon the length of the writing sample (Jewell & Malecki, 2005). These fluency measures include total words written (TWW), words spelled correctly (WSC), and correct writing sequences (CWS). Numerous studies have found these measures to have adequate reliability and significant correlations with criterion measures of written expression (Deno et al., 1982; Espin, Scierka, Skare, & Halverson, 1999; Espin et al., 2000; Gansle et al., 2002; Videen, Deno, & Marston, 1982).

Production independent measures are referred to as measures of writing accuracy, because they do not rely upon the length of the writing sample (Jewell & Malecki, 2005). Production independent measures include percentage of words spelled correctly (%WSC), percentage of correct writing sequences (%CWS), and percentage of legible words (%LW). Tindal and Parker (1989) examined the use of production independent measures, specifically %WSC and %CWS and found these measures to be reliable and valid with middle school students and more strongly correlated with teacher's holistic ratings of student writing than production dependent measures.

Espin et al. (2000) proposed the use of a new scoring measure: correct minus incorrect writing sequences (CMIWS). It was thought that CMIWS, like %CWS would take into account both correct and incorrect writing sequences; however, unlike %CWS, CMIWS would not be limited to a scale of 0 to 100, and thus might be more sensitive to growth. In a study of middle school students, Espin et al. (2000) found CMIWS was reliable and valid for use with 3-minute and 5-minute writing samples.

A review of the research literature on CBM in written expression follows. A brief review of studies completed at the elementary level is followed by a more in-depth review of research completed at the secondary level. For secondary level studies, issues of reliability and validity, more complex scoring measures, type of writing and duration, gender, sensitivity to growth and monitoring progress, and practicality of scoring measures will be reviewed. Then, research completed across grade levels will be presented. Appendix A provides a table of written expression studies reviewed, including grade(s), type of prompt, time, scoring methods, criterion validity, and results and limitations.

Elementary Level Studies

Research on CBM in written expression began with Deno and his colleagues at the Research Institute on Progress Monitoring at the University of Minnesota. Early studies focused on the technical aspects of different scoring measures (Deno et al., 1982; Videen et al., 1982). This review will focus on the three most commonly used scoring measures: TWW, WSC, and CWS.

In the earliest study, Deno, et al. (1982) examined the relationship between TWW, WSC, and performance on different criterion measures. In this study the criterion variables were the Test of Written Language (TOWL) and Developmental Scoring System. Students wrote in response to story prompts, topic sentences and picture stimuli for one to five minutes. Validity coefficients were strongest for three and five minute samples. TWW and WSC were highly correlated with the criterion measures with correlations ranging from .67 to .84. The measures also significantly differentiated

resource room students from general classroom students at various grade levels. Correlations were similar for each type of prompt.

In 1982, Videen et al. introduced the scoring measure CWS. The investigators questioned whether students might begin generating words that would not add meaning to their writing but would improve their writing scores if only TWW and WSC were used to monitor progress. Thus, they suggested that CWS might better reflect improvement but still be an easy scoring measure. Samples from the Deno et al. (1982) study were randomly selected and scored for CWS. Participants were 50 students in third through sixth grades. Results revealed that CWS correlated highly with TWW (r = .93). Correlations between CWS and the Developmental Scoring System were weak (r = .49) and correlations between CWS and the TOWL were moderate (r = .69). Stronger correlations were found between CWS and holistic ratings given by teachers (r = .85)

Marston (1989) and McMaster and Espin (2007) both provided reviews of different types of reliability of scoring measures including test-retest, alternate form, and internal consistency. Most studies also reported interscorer reliability coefficients above .90 for most measures. Marston and Deno (as cited by McMaster & Espin, 2007) reported strong test-retest correlations over a one-day interval and moderate correlations over a three-week interval. Marston and Deno also found strong alternate form reliability between two 5-minute story prompts. Marston reported that alternate form reliabilities ranged from .42 to .95 for TWW and .41 to .95 for WSC. Most reliability coefficients were above .70. Alternate form reliability increased when scores were aggregated across writing samples. No reliability coefficients were reported for CWS.

Marston et al. (as cited by McMaster & Espin, 2007) examined the sensitivity of TWW and WSC for indexing change in student performance. First through sixth grade students wrote samples in the fall, winter, and spring. Results indicated that TWW and WSC demonstrated consistent increases across the year and across grade levels. These results were replicated in a larger scale study completed by Deno, Marston, Mirkin, et al. (as cited by McMaster & Espin).

Tindal & Parker (1991) also examined the criterion validity and sensitivity to growth of different scoring measures. In contrast to early studies, correlations among TWW, WSC, and CWS and analytic scores (1 to 5 on story idea, organization, and mechanics) were only weak to moderate (r = -.02 to .63). Students in grades three through five improved significantly over time on all three measures. Statistically significant differences were found on all measures between students with learning disabilities and general education students and on some measures between students with low performance and general education students, indicating that the measures effectively differentiated among students of different skills levels. Correlations between TWW, WSC, and CWS and the Stanford Achievement Test were in the low to moderate range (r = .18-.41). Correlations with holistic judgments of student writing were strong (r = .85).

In attempt to determine if scoring measures could be used for screening purposes, Parker, Tindal, & Hasbrouck (1991a, Study 1) administered a story prompt in the fall and the spring to students in grades two through five (N = 1,917). All of the students were receiving Chapter 1 or special education services. In the fall and spring, students wrote for up to six minutes in response to a story prompt. Samples were scored using TWW,

WSC, CWS, %WSC, and %CWS. Teachers' holistic ratings served as the criterion variable. Weak to moderate correlations with the criterion variable were obtained at all grade levels with CWS demonstrating the strongest correlation (r = .56). Further analyses examined dispersions in the bottom of the score distributions to determine if the measures could identify students "at risk" for writing difficulties. These analyses indicated that %CWS was the most viable screening tool, because it was moderately correlated with holistic ratings and had a suitable distribution at the lower ranges. However, %CWS was only moderately efficient because it had a 20-percentile point range of uncertainty. The other scoring measures lacked sensitivity to student differences and the investigators cautioned that the use of the measures could lead to false negatives. The investigators noted that most students stopped writing long before the six-minute time limit had passed.

To summarize, research at the elementary level supports the validity of TWW, WSC, and CWS as indicators of students' performance in written expression. In the majority of studies, these scoring measures were reliable and shown to correlate at moderate to strong levels with the criterion. It should be noted that results of the early studies completed by Deno and colleagues produced stronger reliability and criterion correlation coefficients than later studies by Tindal, Parker, and colleagues. Recently, Gansle et al. (2004) found a weak correlation between total words written and the Woodcock Johnson-Revised Writing Samples subtest (r = .23). The scoring measures reviewed discriminated between students in different groups and at a different grade levels, but were not effective in identifying students at risk. More recent studies completed by Gansle and colleagues (2002, 2004) on new scoring measures including

number of nouns, verbs and adjectives, long words, total and correct punctuation, and simple sentences failed to show growth over time, weak to moderate alternate form reliability, and weak criterion validity with standardized achievement tests.

Secondary Level Studies

Research on curriculum-based measurement in written expression with elementary level students was followed by research with secondary level students. However, the majority of secondary level research has focused on middle school students and only one study has focused on high school students (Espin et al., 1999). For secondary students, researchers have examined reliability, validity, sensitivity to growth and monitoring progress, type of writing and duration, gender differences, and practicality of scoring measures.

Reliability and validity. In the first study utilizing a sample of middle school students, Tindal and Parker (1989) applied progress measures to the writing of middle school students in compensatory and special education programs. A sample of 172 students, 30 in special education classes and 142 in "remedial programs," in sixth through eighth grades participated. Students wrote for six minutes in response to a story starter. The samples were scored for TWW, WSC, CWS, legible words (LW), mean length of CWS (ML/CWS), %WSC, %CWS, and %LW. The samples were also scored holistically on a scale of 1 to 7 for communication effectiveness.

Analysis of differences in scoring measures between the special education and remedial program students were conducted (Tindal & Parker, 1989). Significant differences were found between the groups on the holistic rating and on three production independent indices: %CWS, %CSW, and ML/CWS. The intercorrelations of measures

produced two clusters, production dependent and production independent. In general, correlations among scoring measures were in the moderate to strong range. However, correlations between production dependent measures and their production independent counterparts were only low to moderate (CWS and %CSW, r = .36; WSC and %WSC, r = .53; LW and %LW, r = .51). Reliable differences were found between the two student groups on the holistic rating, %WSC, %CWS, and ML/CWS. No reliable differences were found on TWW, WSC, CWS, and LW. To determine the relationship between the eight scoring measures and holistic scores, holistic scores were regressed on each of the eight measures separately. The three production measures were weakly related to holistic scores. In contrast, two production independent measures, %WSC and %CWS, were highly related to holistic scores. Regression of holistic ratings on scoring measures resulted in moderately large coefficients (.59 to .75) for %CWS, %WSC, and ML/CWS. Results indicated that a "production free factor," including WSC, CWS, and LW, was a moderate-to-strong predictor of teacher judgments of communication effectiveness of student writing. The researchers stated that generalizations beyond this sample were limited because it focused on low achieving students and student receiving special education services, and the holistic score judged only the ability to communicate.

A Watkinson and Lee study (1992) found results similar to those of Tindal and Parker (1989). Twenty-six students, in grades six through eight, identified with a learning disability in written language were matched on gender and grade level with randomly selected general education students. All students wrote for three minutes in response to a story starter. The samples were scored for eight measures: TWW, LW, WSC, CWS, number of incorrect word sequences (IWS), %LW, %WSC, and %CWS. Interrater

reliability ranged from .80 for %CWS to .99 for TWW. Significant differences between the two student groups were obtained for two production dependent measures, CWS and IWS, and for all production independent measures, %LW, %WSC, and %CWS. Intercorrelations among scoring indices revealed strong positive relationships among the production dependent measures (r = .776-.949, p < .001). Strong positive relationships were also revealed among the production independent measures (r = .790-.893, p < .001). A moderate positive relationship was found between CWS and the production independent measures (r = .304-.591, p < .05) and IWS showed moderate to strong negative correlations with production independent measures (r = -.765- -.880, p < .001). General education students and students identified with a learning disability significantly differed on CWS, IWS, %LW, %WSC, and %CWS.

In a study conducted Parker et al. (1991a, Study 2), 243 students in grades six (n = 91), eight (n = 89), and eleven (n = 63) wrote for six minutes in response to a story prompt presented in the spring. The samples were scored for TWW, WSC, CWS, %CWS, and %CWS. Holistic ratings of the writing samples communication effectiveness (1 = very poor to 7 = very effective) served as the criterion variable. At each grade level, criterion validity correlation coefficients were in the weak to moderate range for both production dependent measures (TWW, r = .39-.41; WSC, r = .43-.52; CWS, r = .48-.56) and production independent measures (%WSC, r = .34-.46; CWS, r = .36-.42).

Espin et al. (2000) investigated the alternate form reliability and criterion validity of scoring measures with middle school students. Based on previous research findings that the production of correct responses was a predictor of student success, Espin et al. introduced a new scoring measure, correct minus incorrect word sequences (CMIWS).

Espin et al. stated, "The correct minus incorrect score holds the potential for combining the two key features established through previous research – the use of correct and incorrect word sequences and the use of a production measure" (p. 143). Students in sixth, seventh, and eight grades (N = 112) completed four writing samples, two story writing samples and two descriptive writing samples. Students typed their responses. At the end of three minutes, students typed a pound sign, and then continued to write for two more minutes. The criterion variables were the teacher's ratings of student writing proficiency and student performance on a district writing test.

Espin et al. (2000) obtained interscorer agreement ranging from 85% to 92%. The strongest alternate-form reliability coefficients were obtained for TWW, WSC, CWS, CMIWS with reliability coefficients ranging from .72 to .80. They noted that the alternate form reliability coefficients for these measures were within the range reported by Marston (1989) for measures at the elementary level, where the alternate form reliability coefficients ranged from .42 to .96. As evidence for validity, CMIWS scores were compared to teachers' ratings of student writing proficiency and eighth-grade students' scores on a district writing test. Results of a multiple regression analysis found that beyond CMIWS, no other variable added to the strength of the prediction of teacher ratings and performance on the district writing test. Overall, the results suggested that the CMIWS scoring index may be a valid score to use, particularly with middle school students.

More complex scoring measures. In the only study utilizing a sample of high school students, Espin et al. (1999) investigated the use of combining progress measures and using computerized scoring. A sample of 147 tenth-grade students included four skill

groups: students identified with learning disabilities (n = 9), students in a basic English class (n = 39), students in a regular English class (n = 50), and students in an enriched English class (n = 49). During the first 2 weeks of May, students completed a writing sample in response to a story starter. Students had 30 seconds to think and then three minutes to write their story. Based on the hypothesis that as students become older and their writing becomes more complex, Espin et al. sought to determine if a combination of scoring indices best predicted student performance. The investigator typed the samples exactly as written, including all errors, into a word-processing program. Using the grammar check component of the program, samples were scored for TWW, number of characters written, number of characters per word, and number of sentences written. Then, WSC, CWS, and ML/CWS were calculated by hand. The criterion measures included the language subtest of the California Achievement Test (CAT) administered in 11th grade, English grades (first and second semester of 10th grade), and a holistic rating. For the holistic rating, teachers were told to read the writing samples and rate them on a scale of 1 (lowest rating) to 5 (highest rating).

Espin et al. (1999) found significant correlations in the low to moderate range (r = .30-.45) between the criterion measures and CWS, ML/CWS, characters per word, and sentences written. These measures also differentiated students in the four groups. It was noted that the obtained correlations were much lower than those obtained at the elementary level. A combination of measures proved to be the best predictor of student performance, with characters per word, sentences written, and ML/CWS accounting for 38% of the variance in the language arts subtest of the California Achievement Test. Espin et al. concluded that the results imply that each of the variables was tapping into a

different aspect of writing ability and the variables characters per word and sentences written may reflect a more sophisticated level of writing not previously studied at the elementary and middle school levels. The results suggest that a single measure of writing performance may not be enough to capture writing ability at the high school level and it may be necessary to use multiple indicators of proficiency. The study did not investigate the reliability of the measures. The study was limited by the use of CAT results collected in the 11th grade, the school year following the collection of the writing samples completed in the 10th grade.

Type of writing and duration. The technical adequacy of writing sample duration has been studied at the elementary and middle school levels. An early study completed by Deno, Mirkin, and Marston (1980) found that the validity and reliability of writing samples was unrelated to duration at the elementary school level. At the elementary school level, usually a three-minute writing sample is taken (Espin et al., 2000). In later research studies using samples of middle school students, the duration of writing varied from 3 minutes to 10 minutes (Espin et al., 2000; Parker et al., 1991a; Tindal & Parker, 1989). Findings from these studies led Espin et al. (2000) to state, "It may be necessary to collect longer samples of writing to obtain a more accurate representation of students writing skills" (p. 142).

In a first step toward answering this question, Espin et al. (2000) studied the technical adequacy of writing samples of different duration and different types of writing with a sample of middle school students. They reasoned that many secondary level writing is expository as opposed to narrative, so expository writing might better reflect writing proficiency. The purpose of expository writing is to inform the reader or give

facts and information about a topic. In contrast, narrative writing tells a story. Students wrote four writing samples, two expository and two narrative, for three and five minutes. For the expository writings, students wrote in response to the topics "Describe the inside of your school building for someone" and "Describe the clothing that students in your school wear." For the narrative writings, students wrote in response to the story starters "It was a dark and stormy night…" and "I stepped into the time machine…"

Espin et al. (2000) calculated alternate form reliability separately for the story and descriptive writing samples and for the three and five minute writing samples. When scoring for both CWS and CMIWS, similar alternate form reliability correlation coefficients were obtained for TWW, WSC, CWS, and CMIWS across type of writing and duration. Low alternate form reliability correlation coefficients were obtained for words incorrect, characters per word, and ML/CWS. When entered into the regression equation, duration of writing did not contribute to the strength of the prediction of holistic ratings of student writing. While small differences were found in the correlations between CMIWS and scores on a district writing test favoring the longer five-minute sample, the differences were not statistically significant. Overall, the results found few differences in the reliability and validity coefficients across story writing and descriptive writing samples and across three and five minute samples. Espin et al. concluded that the duration of the time samples were still relatively brief, and may not have been long enough to generate meaningful differences for older students.

A more recent study completed by Espin, De La Paz, Scierka, and Roelofs (2005) used a small sample of 22 students in seventh and eighth grades to determine whether the length of text affected reliability and validity and whether CWS and CMIWS were

sensitive to growth. Six students were identified with a learning disability. The remaining students were classified into low (n = 6), average (n = 6), and high (n = 4) achieving groups based on their scores on the written expression subtest of the Wechsler Individual Achievement Test. Students had 35 minutes to write an expository essay, scored using TWW, CWS, and CMIWS. Expository essays were chosen because seventh and eight grade students were required to write expository essays to pass a state mandated assessment. The criterion variables in the study were the number of functional essay elements and holistic ratings of essay quality.

In the beginning of the study, students wrote six essays. After four weeks of writing instruction, students wrote more expository essays. The correlation between CWS and CMIWS in the first 50 words of the writing sample and the criterion variables were calculated to determine the effect of text length. They stated that this type of analysis addressed the issue of whether text length alone was responsible for the correlations between the scoring measures and the criterion measures or whether the scoring measures were also important. Validity correlation coefficients between TWW, CWS, and CMIWS and the criterion measures were moderate to strong (r = .58-.90). Lower validity correlation coefficients were obtained between TWW, CWS, and CMIWS for the first 50 words of each writing sample and the criterion measures (r = .33-.59). The lower validity coefficients may be due to limiting the range of CWS and CMIWS scores.

To study sensitivity to growth over time, a MANOVA with time (pretest to posttest) as a within-subjects factor was run (Espin et al., 2005). Dependent variables entered into the analysis included functional essay elements, quality ratings, TWW, CWS, and CMIWS. Results found that all five scoring measures were sensitive to change

over time. Results of the MANOVA revealed significant effects and follow-up univariate *F* tests revealed significant changes on all five measures. The scoring measures were sensitive to change over time. The number of students in each group was too small to allow for statistical testing, so group differences were inspected for changes. When the length of the text was limited to the first 50 words, low, average, and high achieving writers showed little change over time. In comparison, students identified with a learning disability showed more substantial changes when text length was limited. Based on these findings, Espin et al. concluded that text length is an important factor to consider and students may need more time to write to obtain more reliable and valid samples of writing performance. Espin et al. highlighted the need for future research studies using different time frames and a larger sample size to confirm the findings.

In summary, results from these studies at the secondary level found adequate alternate form reliability and criterion validity for TWW, WSC, CWS, and CMIWS. Also, no differences were found in reliability or validity depending on the type of writing, narrative versus expository. It also appears that older students may need to write for a longer period of time in order to measure growth over time, especially for more proficient writers.

Gender differences. Recent research has found that the gender of the student needs to be considered when deciding what scoring measures will be used to monitor progress (Jewell & Malecki, 2005; Malecki & Jewell, 2003). Mixed results in CBM reading research about the impact of gender (Knoff & Dean, 1994; Kranzler, Miller, & Jordan, 1999) led Malecki and Jewell to investigate the impact of gender in written expression.

Malecki and Jewell (2003) found significant gender differences in performance between first through eighth grade boys and girls. A sample of 946 students provided a three-minute writing sample in response to a story starter in the fall and spring of the school year. The samples were scored using production dependent, production dependent, and accurate production measures. A multivariate analysis of variance was conducted with fall TWW, WSC, CWS, %WSC, %CWS, and CMIWS scores as dependent variables. A gender main effect was found, Wilks' lambda = .923, $F(6, 929) = 12.96$, $p < .001$ (partial $\eta^2 = .077$, very small effect size). Results of follow-up univariate analyses indicated that there were significant differences between boys and girls on all scoring indices, Fs(1, 934) = 49.6, 51.2, 48.7, 18.0, 7.4, and 27,5 respectively, ps < .001 (partial $\eta^2 = .050, .052, .050, .019, .008, .029$, respectively, very small effect sizes). On all indices, girls outperformed boys. In addition, a significant gender x grade level interaction was found for CWS, %WSC, and CMIWS. On CWS, girls' scores were higher than boys' and the gap grew over time. A similar pattern was found for CMIWS. On %CWS, girls outperformed boys in first and second grade, but the gap closed in grades three through five and in middle school.

A follow-up study completed by Jewell and Malecki in 2005 again investigated the impact of gender on written expression measures. A sample of 203 second, fourth, and sixth grade students completed a three-minute writing sample in response to a story starter. The samples were scored for TWW, WSC, CWS, %WSC, %CWS, and CMIWS. Results of a MANOVA found a main effect for gender [Wilks' Lambda = .870, $F(5, 191) = 5.70$, $p < .001$]. Results of follow-up univariate analyses for the gender main effect revealed significant differences only on the production dependent measures (TWW,

WSC, CWS), *F*s(1, 195) = 16.05, 17.96, 10.30, respectively, *p*s < .01. No significant differences were found on the production independent measures or accurate production index. Girls outperformed boys on all of the production dependent measures, writing more words, and producing more correctly spelled words and correct writing sequences. While boys and girls at all grade levels differed in the amount they wrote their writing accuracy was not significantly different. The boys may have been less fluent, but they were equally accurate in their writing.

Malecki and Jewell's findings (2003) led them to caution educators that boys may be over-identified for difficulties in writing if only fluency, production dependent, measures were used and normative data did not take gender into account. Given these differences, separate norms for boys and girls may be needed for fluency indicators. In light of these different findings, the researchers concluded additional research was needed on the impact of gender. In addition, gender differences on writing indices have not been investigated with high school students.

Monitoring progress and sensitivity to growth. Once research had established the reliability and validity of scoring measures, interest turned to their sensitivity to small increments of growth in order measure progress over time (Parker, Tindal, & Hasbrouck, 1991b). The sensitivity of written expression scoring measures for growth monitoring has been discussed by several researchers (Espin et al., 2000; Parker et al., 1991b, Tindal & Parker, 1989; Watkinson & Lee, 1992). The findings of Tindal and Parker (1989) and Watkinson and Lee (1992) that production dependent measures differentiate among students and spreads students out more into a distribution than production independent measures led to the hypothesis that production dependent measures might also be more

sensitive to student growth. Espin et al. (2000) and Tindal and Parker (1989) advocated for the use of fluency measures over "percentage" measures when monitoring student progress.

Tindal and Parker (1989) and Parker et al. (1991a, 1991b) found that %CWS was not appropriate for describing writing growth and cautioned against the use of percentage measures for progress monitoring because percentages may "mask" student growth. While a student's writing fluency may have increased over the course of the academic year, the student's percentage scores may have decreased. For example, a student writes 25 word sequences with 22 correct in the fall and 50 word sequences with 40 correct in the spring. The student increased over the course of the year from writing 22 to 40 correct word sequences; however, the percentage score decreased from 84% to 80%.

Parker et al. (1991b) investigated the utility of scoring measures to measure the progress of special education students. Participants included 36 students in grades six through eight identified with learning disabilities. Students completed 3-minute writing samples four times during the year (October, January, February, and April) in response to story starters. Samples were scored for TWW, WSC, CWS, LW, ML/ CWS, %WSC, and %LW. The criterion variables were holistic judgments of the essay's communication effectiveness and the Test of Written Language (TOWL) administered in May. TOWL scores were correlated with the holistic ratings and the seven measures. Interrater agreement for the holistic ratings ranged from .74-.97 ($p < .001$). Linear growth was found for TWW, WSC, and LW. However, TWW and LW yielded the lowest correlations with holistic ratings and the TOWL. The measures that correlated highly with the TOWL and holistic ratings were %LW, ML/CWS, and CWS. However, a large

amount of variability was obtained for these measures indicating that they may lack the "stability" needed for progress monitoring.

In a later study, Malecki and Jewell (2003) investigated which scoring measures were appropriate for measuring progress with elementary and middle school students. Students in first through eighth grades generated writing samples in the fall and spring of the school year. The writing samples were scored using all three types of scoring indices, production dependent, production independent, and accurate production, to assess possible growth trends.

Overall, older students outperformed younger students on all of the indices (Malecki & Jewell, 2003). At the middle school level, writing fluency and writing accuracy were not closely associated; however, at the younger grades the indices were significantly related. They concluded that at the older grades, accuracy measures (production independent or accurate production) should be considered the most appropriate scores to use. At all grade levels, the measures of writing fluency and the accurate production index increased significantly from fall to spring. In addition, the percentage indices did show significant growth over a span of time for early elementary students. It was noted that the results needed to be interpreted cautiously because the percentage indices may not be as sensitive to student growth in the short-term as production dependent indices. The results of this study contribute evidence that production independent measures may not be as sensitive to growth over time as production dependent measures at the older grade levels. Malecki and Jewell concluded that the sensitivity of the various measures to student progress over time has not been clearly delineated and requires more research.

Results of studies at the secondary level investigating the sensitivity and ability to monitor progress indicate that accuracy measures are the most appropriate scores to use with this age group. Parker et al. (1991b) found some strong criterion validities and growth over the six months of study for some scoring measures, but "while some indices appeared promising in terms of validity, stability, or sensitivity to growth, none was adequate in all of these areas" (p. 79). Based on their findings, Espin et al. (2000, 2005) stated that CMIWS showed the most promise for revealing growth in written expression over time. However, the sensitivity of the measures for measuring progress over time is not clear and requires further research.

Practicality of scoring measures. Another issue that needs to be considered with scoring measures is practicality, specifically the amount of time it takes to score the writing sample. While empirically supported CBM reading measures can be administered and scored quickly, little information is available on the amount of time required to score CBM written expression scoring measures. Information on scoring time is especially important because CBM is supposed to be a quick and efficient task (Gansle et al., 2004). Two recent studies have investigated the amount of time it takes to score writing samples using simple scoring measures.

Malecki and Jewell (2003) stated that they undertook this investigation to provide practical recommendations for how to choose appropriate scoring measures. The amount of time needed to score a writing sample using either TWW or WSC ranged from 22 to 37 seconds, with an average of 30 seconds. As the grade level of the student increased, grades one through eight, the amount of scoring time increased by approximately six seconds between grade levels. Using TWW or WSC, an early elementary student's

writing sample could be scored in under one minute and a middle school student's sample could be scored in just over one minute. The average amount of time needed to score a writing sample using CWS ranged from 46 to 82 seconds, again with the time increasing with grade level. To score a writing sample using all three measures, the average scoring time per sample ranged from 1.5 minutes for an early elementary student to 2.5 minutes for a middle school student.

Gansle et al. (2004) collected data on the amount of time it took to score the writing samples of third and fourth grade students. The average amount of time to score a sample for TWW was 25 seconds with a standard deviation of 12 seconds. The average amount of time to score a writing sample using CWS was 72 seconds with a standard deviation of over 62 seconds.

Simple scoring measures such as TWW or WSC can be scored quickly by hand or via computer, but previous research has found that these measures are not valid for use with secondary students (Espin et al., 1999; Espin et al., 2000; Jewell & Malecki, 2005). Research using samples of middle school students indicates that teachers will have to score student writing by hand in order to gain reliable and valid information on writing (Espin et al., 2000). The amount of time required to score a sample using a more complex measure such as CMIWS will be more time consuming, but research on the amount of time it takes to score a writing sample using more complex measures has not been conducted.

Previous research also indicates that a combination of scoring measures better predicts secondary student performance (Espin et al., 1999). However, while a combination of scoring measures may improve usefulness it may reduce practicality for

classroom use (Espin et al., 1999). “Every measure that is added to the system makes the system more complicated and the more complicated the system, the less likely it is to be used and maintained over an extended period of time” (Espin et al., 1999, p. 19). Given the fact that high school teachers teach a large number of students throughout the school day, research on the amount of time it takes to score longer writing samples using more complex scoring measures is needed.

Summary of secondary level studies. Results of studies completed at the middle level found that scoring measures reliably discriminate among student groups (Tindal & Parker, 1989, Parker et al., 1991b). Studies have found adequate alternate form reliability and criterion validity for TWW, WSC, CWS, and CMIWS. Also, there were not differences in reliability and validity with different types of writing. Research on the impact of writing duration suggest that older students may need more time to write depending on the students’ level of writing proficiency. Studies of gender differences in written expression have produced differing results. One study found that boys outperformed boys on production independent, production dependent, and accurate production measures, but a later study found only significant differences on production dependent measures. Discussions by researchers on which measures are most appropriate to use to monitor growth over time have advocated for the use of production dependent measures because production independent measures may not be as sensitive and may mask student progress. Two studies investigated the amount of time it takes to score a sample using simple measures, but research is lacking on the amount of time it would take to score a secondary level student’s lengthier writing using more complex measures such as CMIWS . Overall, the majority of the studies focused on middle school students

and more research is needed on the technical adequacy of measures with high school students.

Studies Across Grade Levels

The research findings presented thus far suggest that certain types of writing measures are more appropriate for use with students of certain ages and that the relationship between scoring measures and scores on other writing criterion, such as published standardized tests, may change with age (Espin et al., 2000; Jewell & Malecki, 2005; Tindal & Parker, 1989). Research at the elementary level found that simple scoring measures such as TWW or WSC were reliable and valid (Deno et al., 1982; Videen et al., 1982). However, research at the secondary level found stronger reliability and validity for percentage measures (Tindal & Parker, 1989) and more complex scoring measures such as CWS and CMIWS (Espin et al., 1999; Espin et al., 2000; Fewster & MacMillan, 2002; Parker et al., 1991a, 1991b; Tindal & Parker, 1989; Watkinson & Lee, 1992). The contradictory findings led some researchers to investigate the technical adequacy and utility of scoring measures across grade levels.

Malecki and Jewell (2003) investigated production dependent (TWW, WSC, CWS), production independent (%WSC, %CWS), and accurate production (CMIWS) measures across first through eighth grades. In the fall and spring of the school year, 946 students provided 3-minute writing samples in response to a story starter. The interrelationships between the scoring measures were calculated through correlational analyses. Most of the scores were highly related to one another, except the production independent measures (%WSC and %CWS) were not significantly related to TWW at the middle school level. With older students, how much students wrote was not closely

associated to the accuracy of their writing. At all grade levels, the CMIWS scores related well with both the production dependent and production independent measures leading Malecki and Jewell to conclude that the CMIWS measure is tapping aspects of both fluency and quality.

A series of six repeated measures Multivariate analyses of variance (MANOVA) were conducted to determine if differences existed between fall and spring writing scores by grade level (Malecki & Jewell, 2003). Results found significant differences between fall and spring scores on all fluency measures including TWW (Wilks' Lambda = .866, $F(1, 809) = 124.79$, $p < .001$), WSC (Wilks' Lambda = .860, $F(1,809) = 131.23$, $p < .001$) and CWS (Wilks' Lambda = .854, $F(1, 809) = 138.16$, $p < .001$). At all graded levels and for all production dependent indices, students' scores were higher in the spring than in the fall (partial η^2s = .134, .140, 146, respectively, small effect sizes). In addition, CMIWS scores were significantly higher from fall to spring for all grade levels (Wilks' Lambda = .887, $F(1, 809) = 103.33$, $p < .001$, partial $\eta^2 = .113$, small effect size. A significant grade level interaction was present for %WSC (Wilks' Lambda = .939, $F(2, 809) = 26.23$, $p < .001$) and for %CWS (Wilks' Lambda = .884, $F(2, 809) = 53.03$, $p < .001$). In both cases there was a significant difference between fall and spring scores only for first and second grade students (partial η^2s = .061 and .116, respectively, very small effect sizes). No significant differences were found over time on the percentage scores for grades three through eight.

In a follow-up study, Jewell and Malecki (2005) continued to research the utility of the three written expression scoring measures and compared scores to both direct and indirect criterion measures of writing ability across a sample of 203 second-, fourth-, and

sixth-grade students. The criterion measures included utilized were a curriculum-based measure of written expression, the Tindal and Hasbrouck analytic scoring system (THASS), the Stanford Achievement Test (SAT), and students' Language Arts grades from the fall semester. The THASS was chosen for inclusion in the study because it was similar to statewide standardized writing assessment techniques. One 3-minute writing sample was collected from students in response to a story starter. The samples were then scored six different ways (TWW, WSC, CWS, %WSC, %CWS, and CMIWS). The interrelationships between the scoring measures were calculated through correlational analyses. Most of the scores were highly related to one another except the production independent measures were not significantly related to TWW at any grade level. For sixth-grade students, WSC was not significantly related to the production independent measures. This led Jewell and Malecki to conclude that with older students the WSC was not related to measures of writing accuracy. CMIWS scores did relate well with both the production dependent and production independent indices measures.

Results found grade level differences in how measures of written expression related to students' scores on the criterion measures. With older students, production independent and accurate production measures were more related to standardized achievement scores, an analytic rating, and grades than measures of writing fluency. The correlations between the production independent and accurate production measures for all grade levels were significantly related to the SAT language subtest scores (r = .34 to .67, $p < .01$) and the THASS scores (r = .34 to .58, $p < .01$). While significant, these correlations are in the weak to moderate range. The investigators concluded that the CMIWS scores do seem to be tapping aspects of both writing fluency and accuracy, as

evidenced by significant correlations with the production dependent and production independent scoring measures and the criteria. For sixth grade students, CWS continued to be significantly related to the criterion, but other production dependent measures were not significantly related to the criterion measures at the older grade levels. It appears that examining only the quantity of these students' writing is not assessing the skills being measured by the criterion measures. At all grade levels, measures of writing accuracy may be more strongly related to students' performance on other types of writing criteria than measures of writing fluency. It is important to note that the results of this study were limited by the use of only one scorer so interscorer reliabilities were not calculated.

Jewell and Malecki (2005) concluded that simple scoring measures such as TWW and WSC become less valid as grade level increases and suggested using percentage measures or CMIWS for secondary students. In addition, their findings were consistent with those of Tindal and Parker (1989) who found that production independent scoring measures were more closely related to teachers' holistic ratings than production dependent scoring measures. Jewell and Malecki concluded that the results of the study add to the understanding of writing scoring measures, and help to emphasize that the assessment that most closely relates to the skills needing to be assessed and that is most valid for the intended purpose should always be used.

Weissenburger and Espin (2005) were the first to examine the technical adequacy of CBM writing scoring measures across three different grade levels and its criterion related validity with a statewide assessment, the Wisconsin Knowledge and Concept Examinations (WKCE). A sample of 484 students in 4th-, 8th-, and 10th-grades completed two writing samples in response to story starters within a two-week period

prior to and following the administration of the WKCE. The WKCE is derived from the TerraNova Assessment Series and the CTB Writing Assessment System. The Normal Curve Equivalent (NCE) scores for the WKCE Language Arts subtest and the holistic writing scores from the WKCE (i.e., CTB Writing Assessment System) were used as criterion measures. It was noted that the WKCE Writing Assessment was not administered to 10th-grade students due to Wisconsin's effort to pilot test items that year for a proposed statewide graduation test. Only 4th and 8th grade holistic writing scores were obtained for the study. The writing samples utilized were two stories written in response to story starters scored in 3, 5, and 10-minute segments for TWW, CWS, and CMIWS for each sample length.

First, alternate form reliability was investigated through correlations between the scores obtained from the two story starters for each sample length. Significant alternate form correlation coefficients were found across all three grade levels (r = .55-.84, p < .001). Alternate form reliabilities increased with an increase in sample duration across all scoring methods and grade levels. For grades 8 and 10, only CMIWS for 10 minutes yielded reliability coefficients above .80 (r = .82, .80, respectively, p < .001). While the alternate form reliability coefficients decreased by grade level, the trend was less prominent for the more complex measure of CMIWS and for scores derived from longer writing samples. Then, the criterion validity of CBM scores and WKCE Language Arts scores were investigated. Correlations coefficients for 18 out of 27 grade level correlations reached significance (r = .26-.69, p < .001). The correlations were stronger for CWS and CMIWS than for TWW at all grade levels. Relative to 4th and 8th grades, 10th grade correlations were low and only a few reached statistical significance.

Correlation coefficients between CBM scores and WKCE Writing Assessment holistic scores revealed correlation coefficients ranging from .33 to .65 ($p < .001$). For both 4th and 8th grade, the strongest correlations were between CMIWS and holistic scores (r = .56 and .65, respectively).

While significant positive relations between most measures and the WKCE tests were found, only the moderate to large correlation coefficients for the 4th- and 8th-grade students for CWS and CMIWS were strong enough to provide sufficient evidence to substantiate the validity of these measures at these grade levels. At all grade levels, CMIWS was the strongest predictor of performance, CWS was the second strongest, and TWW was the weakest. Based on the results, the authors concluded that the results of the study do not support the validity of any scoring method at the 10th grade level. The results of the study were limited by several factors including the use of only students in Wisconsin and a sample that was 96% Caucasian. Holistic writing scores could not be obtained for the 10th-grade students and the authors noted that it was possible CBM of writing may have been more strongly related to the holistic writing scores at the 10th grade level.

In summary, findings from the three studies investigating scoring measures across grade levels suggest that the criterion validity of CBM decreases as students get older. While the validity of measures administered in the Weissenburger and Espin (2005) study did not increase substantially with time, previous research has indicated that longer samples do increase the validity of writing scores (Espin et al., 2005). Therefore, the effect of sample duration on the criterion validity of samples needs further study (McMaster & Espin, 2007).

Summary of Research Findings

The implementation of high-stakes tests required for a student to graduate from high school creates a need to provide high school teachers with research-based progress measures of writing that are efficient, reliable, and trustworthy (Fuchs, 2004). Information gained from these measures could be used to monitor student progress in writing and guide instructional practice. However, few technically adequate measurement systems are available for high school teachers (Fewster & MacMillan, 2002). CBM is a research validated method for measuring student growth and progress in many academic skills areas (Deno, 1985). While there is a broad research base providing evidence of the technical adequacy of CBM with elementary and middle school students, less attention has been paid to the technical adequacy of CBM for use with high school students. While CBM of writing has been empirically validated at the elementary level, CBM research at the secondary level is limited (Fewster & Macmillan, 2002). Further research related to CBM writing measures has been advocated by multiple researchers (Espin et al., 2004; Espin et al., 2005; Fewster & MacMillan, 2002; Malecki & Jewell, 2003; Weissenburger & Espin, 2005).

Three different kinds of writing measures, production dependent, production independent, and accurate production, have demonstrated adequate reliability and validity for use with elementary and middle school students (Deno et al., 1980; Espin et al., 2000; Jewell & Malecki, 2005; Tindal & Parker, 1989). Other studies have provided evidence of the criterion validity of writing scoring measures for elementary and middle school students on standardized and high stakes assessments, but this type of research has not been completed with high school students.

Three other factors have been studied in the previous literature: length of sample, gender differences, and time needed to score samples. Espin et al. (2005) reported preliminary results indicating that the length of the writing sample needs to be considered and the older students may need to write for a longer period of time to obtain reliable and valid samples of written expression performance. Results have also indicated that a longer sample may be needed to produce reliable measures of growth in student writing across the school year (Espin et al., 2005; Parker et al., 1991b). Recent findings of significant gender differences in writing performance among elementary and middle school students suggest that girls at all grade levels write more than boys, but their writing accuracy is not significantly different (Jewell & Malecki, 2005; Malecki & Jewell, 2003). Finally, information on the amount of time needed to score writing samples has been reported as ranging from 1.5 minutes for an early elementary student to 2.5 minutes for a middle school student (Malecki & Jewell, 2003). Research on the amount of time it takes to score writing samples using the more complex measures found to be reliable and valid for use with secondary students has not been completed and will be critical in determining its practicality for classroom implementation (Espin et al., 1999).

Studies across grade levels provided evidence that certain types of CBM writing measures are more appropriate for use with students of certain ages and that the relationship between CBM scores and scores on other writing criterion, such as published standardized tests, may change with age (Espin et al., 2000; Jewell & Malecki, 2005; Tindal & Parker, 1989).

In summary, the research findings presented indicate that there is evidence for the reliability and validity of all three categories of writing scoring measures for use with elementary and middle school students; however, the research on CBM in writing with high school students is just emerging and there is not clear support for their validity and reliability with this age group (Espin, Weissenberger, & Benson, 2004). Findings from studies of middle school students indicate that CMIWS may be the best indicator of student writing ability (Espin et al., 2000; Tindal & Parker, 1989; Watkinson & Lee, 1992). Jewell and Malecki (2005) stated, "The task that remains is to determine for what specific use(s) and for whom each index is suitable" (p. 29). Research results indicate that a particular writing scoring index may be more appropriate to use for certain assessment purposes or with students of different ages and gender (Jewell & Malecki, 2005; Malecki & Jewell, 2003). There is a clear need for more research on written expression scoring measures in several different areas including reliability, validity, duration of writing samples, and utility and adequacy as a progress measure.

Research Questions

The purpose of this study is to evaluate the potential utility of writing samples collected by classroom teachers to monitor student progress and predict scores on a high stakes assessment. The study will build on previous research completed at the elementary and middle school level on written expression scoring measures and examine the utility of these measures, developed in the CBM research literature, with high school students as an indicator of overall writing skill. Research has shown that the following issues need to be addressed: the purpose of the assessment (Jewell & Malecki, 2005), grade level (Espin et al., 1999, 2000; Jewell & Malecki; Tindal & Parker, 1989), duration (Espin et al.,

2000; Espin et al., 2005; Parker et al., 1991b), gender (Jewell & Malecki; Knoff & Dean, 1994; Kranzler et al., 1999; Malecki & Jewell, 2003), and practicality and time (Malecki & Jewell). In response to the limitations noted by Espin et al. (2000), this study will address the validity and reliability of written expression scoring measures for measuring progress over time using a sample of high school students.

Although typically CBM research has used timed measures, not all curriculum-based measurement techniques have been timed. The use of untimed writing samples will add to the research on the effect of sample text length, especially important for high school students whose samples tend to be longer (Espin et al., 2005). There are two ways in which the use of untimed measures increases ecological validity. The samples are based on actual classroom practices using teacher-directed assignments and pacing. In addition, use of untimed writing samples will also produce samples that are more ecologically valid due to similarity with the written responses the student is required to produce on the criterion variable used in this study, the English 2 Maryland High School Assessment. The brief constructed response and extended constructed response writing items on the assessment do not have a time limit and the student has a total of two and a half hours to complete all items on the assessment.

Recent research findings of gender differences in writing at the elementary and middle school level will be explored to determine if these differences are also found at the high school level. It would be expected that significant gender differences in written expression will be found. Significant gender differences would be expected on production dependent measures but no significant gender differences would be found on production independent measures.

In addition, this study will address whether different written expression scoring measures have differential predictive validity with the high stakes English 2 assessment required for graduation. The scoring measures and indices to be used here, drawn from the CBM literature, include TWW, WSC, CWS, CMIWS, %WSC, %CWS, production dependent index, and production independent index.

The issue of practicality and time needed to score writing samples using the more complex measures which appear to be the best indicator of secondary student writing proficiency, will also be addressed. This is a critical issue given that the average high school teacher may have over 100 students. If a scoring measure is reliable and valid, but takes a lengthy amount of time to score then it is unlikely to be implemented by classroom teachers.

The study is designed to answer the following questions:

1. What is the reliability of written expression scoring measures and indices for use with 10th-grade students?

    a. What is the alternate form reliability of written expression scoring measures and indices between three BCRs?

    b. Does the reliability of written expression scoring measures and indices differ depending on the length of the text scored?

2. Are there significant gender differences in 10th-grade students' writings?

    a. Are there significant gender differences on production dependent measures?

    b. Are there significant gender differences on production independent measures?

3. What written expression scoring measures and indices are sensitive to growth in 10th-grade students' writing?

4. What written expression scoring measures and indices are valid for predicting 10th-grade student performance on a state mandated high stakes assessment?

Methods

The purpose of this study is to investigate the reliability and validity of written expression scoring measures and indices for predicting success on a high stakes assessment mandated for 10th-grade students. In addition, the sensitivity to measure progress over time, gender differences in writing, and amount of time needed to score the writing samples were also examined. Samples of 10th-grade students writing from the fall, winter, and spring were scored eight different ways. Table 1 includes the scoring methods and how they were computed.

Participants and Setting

Participants in the study attended a large high school in a suburban Maryland school system. The school had a total enrollment of approximately 1400 students. In the school, 18.9% of students qualified for free/reduced lunch, 3.4% were identified as limited English proficient (LEP), and 8.9% received special education services. Approximately 45.6% of the enrolled students were Caucasian, 37% African or African American, 9.8% Asian, 6.7% Hispanic, and .4% Native American. Table 2 presents the passing rates on the English 2 HSA for the sample, school, local district, and state.

Participants were selected from a 10th grade English class in the school where the study took place. Tenth grade students were chosen for this study because they are required to pass the English 2 High School Assessment (HSA) in order to graduate with a

Table 1

Scoring Measures, Definitions, and Computation Method

Scoring Measures	Definition	Computation
Production Dependent		
Total Words Written (TWW)	A count of the total number of words written.	TWW
Words Spelled Correctly (WSC)	A count of the number of words that are spelled correctly. A word is spelled correctly if it can stand alone as a word in the English language.	WSC
Correct Writing Sequences (CWS)	A count of the correct writing sequences found in the sample. A correct writing sequence is defined as two adjacent writing units that are acceptable within the context of what is written. Correct writing sequences take into account correct spelling, grammar, punctuation, capitalization, syntax, and semantics.	CWS
Production Independent		
Percentage of Words Spelled Correctly (%WSC)	The percentage of words spelled correctly in the sample.	WSC/TWW
Percentage of Correct Writing Sequences (%CWS)	The percentage of correct writing sequences in the sample.	CWS/CWS + Incorrect Writing Sequences
Accurate Production		
Correct Minus Incorrect Writing Sequences (CMIWS)	The number of correct writing sequences minus the number of incorrect writing sequences.	CWS – Incorrect Writing Sequences

Table 2

HSA Passing Rate % by Sample, School, District, and State

Level	Year 1	Year 2
	All Students	
Sample	43.2	64.9
School	68.7	83.6
District	78.2	85.7
State	60.1	70.9
	Males	
Sample	44.0	57.9
School	58.7	81.2
District	73.0	82.8
State	51.9	66.0
	Females	
Sample	42.1	72.2
School	78.2	85.5
District	83.6	88.6
State	68.2	75.7

high school diploma. The English 10 essential curriculum is composed of four units: The World of Romance, The Tragic Stance, Satire: The Pen as Scalpel, and The Search for Self. In each unit, students read literature that relates to the theme, determine characteristics of the theme, and compose essays.

Three English 10 teachers in the school where the study took place provided the investigator with their classes' cumulative writing folders. A review of the writing folders found that each classroom teacher had assigned different writing topics. In order to compare student writings on the same topics, it was determined that only student writings from one of the classroom teachers would be utilized in this study. The classroom teacher with the highest number of writing samples was chosen for the study.

Overall, 158 cumulative writing folders from one classroom teacher were reviewed and 97 met the criteria for inclusion in the study. Students were included in the study if their cumulative writing folder contained three brief constructed responses (BCRs) and one extended constructed response (ECR), or two ECRs. Table 3 presents data from each year of the study including the year, topic, and number of writing samples.

Over the two years of data collection, 97 students, including 50 males and 47 females, met the criteria for inclusion. All of the students were enrolled in a general education or honors level English class, but course level was not available for the participating students. Twelve students in the sample, 5 males and 7 females, were receiving special education services. Of the 12 student receiving special education services, 5 were identified with a Specific Learning Disability, 4 were identified with a Speech Language Impairment, 2 were identified with an Emotional Disturbance, and 1 was identified with Other Health Impaired. Nine students, 3 males and 6 females, were identified as Limited English Proficient (LEP). Forty-five students were African American, 30 students were Caucasian, 7 students were Asian, and 7 students were Hispanic. The ethnicity of 8 students was unknown. The number of participants eligible

Table 3

Writing Sample Counts by Year

Sample	Year 1	Year 2
The Stone Boy	45	41
Oedipus 1	37	40
Oedipus 2	37	40
Oedipus 3	37	40
Lord of the Flies	44	36

for free/reduced lunch was unavailable. Table 4 presents student demographic information by year of study.

Predictor Variables

The predictor variables in this study were scores on 10th-grade students' writing samples collected over two academic years. Writing samples were scored using six different written expression scoring measures. In addition, two combinations of scoring measures, a production dependent index and a production independent index, were also calculated. Separate scores were calculated for each writing sample by year, by gender, and for the entire sample.

Production dependent measures. Three production dependent measures were scored: total words written (TWW), words spelled correctly (WSC), and correct writing sequences (CWS). The definitions provided by Jewell and Malecki (2005) were used in this study. Total words written was a count of the total number of word units written in the sample, regardless of spelling or usage. "A word is defined as any letter or group of letters separated by a space, even if the word is misspelled or is a nonsense word" (p. 32).

Table 4

Gender and Ethnicity Counts by Year

Category	Year 1	Year 2
Gender		
Female	24	24
Male	26	23
Ethnicity		
African American	25	20
Asian	3	4
Caucasian	14	16
Hispanic	4	3
Unknown	3	5

Words spelled correctly was the total number of correctly spelled words in the sample, regardless of appropriate usage. “A word is spelled correctly if it can stand alone as a word in the English language” (p. 32). Correct writing sequences was the number of sequences between two adjacent writing units. “A correct writing sequence is defined as two adjacent writing units (i.e., word-word or word-punctuation) that are acceptable within the context of what is written. Correct writing sequences take into account correct spelling, grammar, punctuation, capitalization, syntax, and semantics” (p. 32). A production dependent index was calculated as the sum of TWW, WSC, and CWS.

Production independent measures. Two production independent measures were scored: percentage of words spelled correctly (%WSC) and percentage of correct writing sequences (%CWS). Percentage of words spelled correctly was the percentage of words

in the sample that were spelled correctly. It was calculated by dividing the number of words spelled correctly by the total number of words written. Percentage of correct writing sequences was the percentage of correct writing sequences in the sample. It was calculated by dividing the total number of correct writing sequences by the total number of possible writing sequences in the sample. A production independent index was calculated as the sum of the percentage of words spelled correctly and the percentage of correct writing sequences.

Accurate production measure. An accurate production measure was scored: correct minus incorrect writing sequences (CMIWS). This measure was calculated by subtracting the number of incorrect writing sequences from the total number of correct writing sequences in the writing sample (Espin et al., 2000). Two adjacent words or writing sequences were scored as an incorrect writing sequence when one or both units were syntactically incorrect, grammatically incorrect, incorrectly spelled, incorrectly capitalized, or incorrectly punctuated (Jewell & Malecki, 2005). Espin et al. (2000) also refers to correct minus incorrect writing sequences as an accurate production index.

Criterion Variable

The criterion variable in the study was performance on the Maryland High School Assessment (HSA) in English 2. The Maryland HSAs are high-stakes tests required for graduation with a high school diploma and are used for the purposes of meeting the No Child Left Behind Act requirements (Maryland State Department of Education, 2007). The HSAs, developed as end-of-course exams, are designed to assess students' knowledge and mastery of the Core Learning Goals for the subject areas of English 2, algebra/data analysis, government, and biology. The HSAs are referred to as

"end-of-course" assessments because students take each assessment as they complete the appropriate courses.

Core Learning Goals. The English 2 HSA is a test of knowledge of the Core Learning Goals contained in the English course students complete in 10th grade (School Improvement in Maryland, n.d.). Each of the four Core Learning Goals are further defined by student expectations. The assessment tests student knowledge of the Core Learning Goals at the indicator level and some indicators have assessment limits which indicate more specifically what is assessed.

Goal 1 states that the student will be able to demonstrate the ability to respond to a text by using personal experiences and critical analysis. Goal 2 states the student will be able to demonstrate the ability to compose in a variety of modes by developing content, employing specific forms, and selecting language appropriate for the particular audience and purpose. Goal 3 states that the student will be able to demonstrate the ability to control language by applying the conventions of Standard English in writing and speaking. Goal 4 states that the student will demonstrate the ability to evaluate the content, organization, and language of texts.

Item responses and scoring. The English 2 HSA includes a combination of three different types of test items: selected responses (SR), brief constructed responses (BCR), and extended constructed responses (ECR) (Maryland State Department of Education, n.d.). The assessment covers approximately 60% of course content and consists of 46 SR items, 2 BCRs, and 2 ECRs.

Selected response items are multiple-choice items asking the student to discriminate among a variety of alternatives and to identify the most appropriate

alternative in response to the question (Maryland State Department of Education, n.d.). A response to this type of question usually takes one minute. All SR items are machine scored.

Brief constructed response (BCR) items are open-ended items requiring the student to write an answer consisting of a few sentences with the opportunity to generate and weave ideas into a short response (Maryland State Department of Education, n.d.). It is estimated that the average student will require 8 minutes to answer the question. BCR items are hand scored using a process known as "modified holistic scoring." The amount of credit awarded to the response depends upon the outcomes measured by the item. A generic rubric was developed for scoring the BCR responses on a scale of 0 to 3 in which a score of 3 is the highest possible (see Appendix B for the BCR rubric).

Extended constructed response (ECR) items are open-ended and complex essay items that require the student to produce a response in the form of a paragraph (Maryland State Department of Education, n.d.). It is estimated that the average student will need 15 minutes to complete an ECR item depending on the complexity of the question. ECR items are also hand scored using modified holistic scoring and the amount of credit awarded depends upon the outcomes the item is designed to measure. A generic rubric was developed for scoring the ECR responses on a scale of 0 to 4 in which a score of 4 is the highest possible (see Appendix C for the ECR rubric). Appendix D provides a description of the process involved in selecting and training scorers and scoring the responses.

Validity. Validity is one of the most important aspects of an assessment and refers to the degree to which the assessment is aligned with the content it is intended to measure

(Maryland State Department of Education, 2007). The process of establishing validity begins with the test design and continues throughout the entire assessment process, including design, content specifications, item development, psychometric quality, and inferences made from test results. A content expert with knowledge and teaching experience related to the course in which the HSA was to be administered oversaw the development of test content for all four HSAs. Appropriate content leads who had similar qualifications reviewed the test development work of these individuals.

During the ETS test development process, MSDE had opportunities to review test content and make changes to ensure that the items were valid measures of the knowledge and skills of Maryland students according to course standards (Maryland State Department of Education, 2007). Every item created was referenced to a particular instructional standard (i.e., goal, expectation, or indicator). In addition, the specific reference was confirmed or changed to reflect changes to the item. When the item went to a committee of Maryland educators for a content review, the members of the committee made independent judgments about the match of the item content to the standard it was intended to measure, and evaluated the appropriateness for the age of students being tested. These judgments were tabulated and reviewed by the content experts who used the information to decide which items would advance to the next stage of development.

According to the Maryland State Department of Education (2005), the development of test items involved the collaboration of teachers and other educators, parents, business leaders, community members, and educational and professional organizations from across Maryland. Each test item undergoes a comprehensive review process before inclusion on an assessment. The review process analyzes the content,

style, and language of all items. Revision of items is often necessary to verify content, improve the item stem and/or answer choices, to make the item clear and concise, ensure that it appropriately addresses the indicator it was designed to measure, and to make the item accessible to all students. Staff from MSDE and the Educational Testing Service, the HSA test development contractor, review test items several times. Then, an independent Item Review Team made up of content experts who were not part of the development process reviews the items. Finally, items receive a comprehensive Sensitivity Review to detect any form of bias that may have been overlooked during the other reviews.

The SR item type tests a wide range of knowledge, application, and reasoning skills. The most common type of SR offers the student four answer choices. The typical SR item consists of a stem phrased as a question or an incomplete statement and response options consisting of the correct response and distracters. Detailed item writing guidelines dictate format, content, structure, and response development including general, correct option, and distracters.

The BCR and ECR item types measures the student's ability to analyze and respond to complex situations and text. The item may be presented as a paragraph of prose or a display of visual and/or verbal material. The student supplies a response in the form of a few sentences for a BCR or a lengthier response for an ECR. These items measure achievement in relation to one Core Learning Goal indicator. Specific guidelines and a rubric worksheet guide the development of BCR and ECR items before writing an actual item. Appendix E provides a detailed description of the BCR and ECR item writing guidelines and sample BCR and ECR items from a 2007 publicly released HSA (available at http://hsaexam.org/sample/english.html).

The Core Learning Goals detail the constructs measured by each HSA (Maryland State Department of Education, 2007). All ETS content staff working on item development participated in training on the Core Learning Goals. Test "blueprint documents," created in collaboration with committees of Maryland educators, were developed from the Maryland goals, expectation, and indicators. No other validity data are reported for the English 2 HSA.

Scores. The Maryland State Board of Education set the passing score for the English 2 HSA in October 2005 (Maryland State Department of Education, 2007). The scores were set after conducting thorough standard-setting activities involving more than 100 teachers, administrators, instructional supervisors, parents, and testing experts. The English 2 HSA fulfills the high school testing requirements of NCLB, requiring states to report the number of students performing at proficient and advanced levels.

Maryland uses a scale score to provide a more precise measurement of student achievement and to assure that tests given at different times are comparable (Maryland State Department of Education, 2007). The HSA reporting scale ranges from 240 to 650, and the scores have a mean of 400 and a standard deviation of 40. The scores represent ability estimates obtained using Item Response Theory. Scale scores based on maximum likelihood estimates are reported for the total test scores. While the total test score is based on item-pattern scoring, the subscores are based on raw score to scale score scoring tables. The scale scores also correspond to categories: Basic, Proficient, and Advanced. Scores in the Basic range of 386 to 395 are not passing. The cutoff score for the Proficient range is 396 and is the lowest possible passing score. The cutoff score for the

Advanced range is 429. Analyses for this study were completed using both the categorical and scale scores.

Approximately nine weeks after taking the assessment, the scoring company sends students' scores to the local school system (Maryland State Department of Education, 2005). The local school system then sends students' scores to parents. Parents receive the individual test scores for their student along with scores for the school, local school system, and state. Scores are reported annually in mid-August on the Web at http://www.mdreportcard.org. The state requires that local school systems print HSA scores on all official high school transcripts.

Reliability. The Maryland High School Assessment Technical Report defines reliability as "the extent to which differences in scores reflect true differences in knowledge, ability, or skill being tested rather than fluctuations due to chance or other factors not being tested" (Maryland State Department of Education, 2007, p. 22). Variance in the distribution of test scores is partly due to real differences in knowledge, ability, or skill being tested (true-score variance) and partly due to random error in the measurement process (error variance). Internal-consistency reliability estimates obtained from analysis of the consistency of the performance of students on items within a test were reported. The English 2 HSA contains mixed item types, so it was determined that it was more appropriate to report stratified alpha. "Stratified alpha is a weighted average of Cronbach's alpha for item sets with different maximum score points" (Maryland State Department of Education, 2007). Reliability analyses of the total test scores indicated that all of the HSAs were highly reliable with reliabilities ranging from 0.89 to 0.95 for the primary forms, and from 0.90 to 0.96 for the make-up forms.

The accuracy of decisions based on specified cut scores was assessed for Reliability of Classification using a proprietary ETS computer program RELCLASS. RELCLASS provides two statistics that describe the reliability of classification based on test scores. More specifically, information from an administration of one form is used to estimate decision accuracy and decision consistency. Decision accuracy describes the extent to which students are classified the same way based on the average of all possible forms of the test. Decision consistency describes the extent to which examinees are classified in the same way as they would be on the basis of a single form of a test other than the one for which data are available.

For the January, May, and July 2006 English 2 HSA administrations 51.5% of students achieved scores in the Proficient or Advanced range (Maryland State Department of Education, 2007). More specifically, 48.5% of students scored in the Basic range, 34.1% in the Proficient range, and 17.4% in the Advanced range. Results found decision accuracy values are .81 and .83 for all classifications and .90 and .91 for the Proficient and Advanced classifications, for the January and May 2006 tests, respectively. Therefore, the agreement between classifications based on an observable variable (scores on one form of a test) and classifications based on an unobservable variable (the test takers' true scores) was rated as "very good." Decision consistency values were .73 and .76 for all classifications and .87 and .90 for the Proficient and Above for the January and May tests respectively. Since decision consistency statistics describe the agreement between classifications based on two variables (scores on the form students have taken and a parallel form of the same test that is not administered to the students, these values were within the acceptable range.

Test administration. The English 2 HSA is administered in one school day in January, May, and July and takes approximately three and a half hours to complete (Maryland State Department of Education, n.d.). The first and second testing sessions last 60 minutes and each session is followed by a 5-minute break. The third testing session lasts 50 minutes. Testing accommodations are provided as designated in a student's Individualized Education Plan or Section 504 plan. Students absent during regular HSA testing must take the test on one of the scheduled make-up days. The make-up schedule is set at the same time as the regular testing schedule.

Data Collection Procedures

The English 10 classroom teacher whose students writing samples were used in this study provided the following information regarding the development of the writing prompts and directions given to the students. Two English 10 teachers wrote the prompt for *Lord of the Flies* and "The Stone Boy" ECRs. Another English 10 teacher wrote the prompt for *Oedipus* BCRs. Appendix F provides "The Stone Boy" ECR. Appendix G provides the three *Oedipus* BCR writing prompts. Appendix H provides the *Lord of the Flies* ECR. Students wrote "The Stone Boy" ECR in September, *Oedipus* BCRs in November and December, and *Lord of the Flies* ECR in March. All students were directed to place the graded BCRs and ECRs in their cumulative writing folders; however, the teacher did not monitor this. During the school year, the teacher stated that she does modify her instruction based on her evaluation of the student's writing. In addition, the students review good examples of BCRs and ECRs and edit practice paragraphs. The teacher used different methods for assigning grades and giving students feedback on their writing. The teacher sometimes scored the paper according to the HSA

rubric or assigned a letter grade and points (0-100) according to a teacher developed rubric. The teacher also provided written feedback in the form of comments on their papers.

Prior to writing "The Stone Boy" ECR, students read the short story and analyzed the characters and themes through class discussion. After the discussion, the students took a quiz on the story. During another class period, students analyzed the poem "Weapons" by McGinley as a warm-up exercise. Then, the class discussed how the poem related to the short story. The teacher distributed "The Stone Boy" prompt and read it aloud. She wrote the thesis statement with the class and "walked them through" an outline of the paper. The students completed the ECR for homework.

Students wrote the *Oedipus* BCRs over three consecutive weeks. Prior to writing the Oedipus 1 BCR, students completed a warm-up exercise where they defined pride in their own words and wrote on real life example of it. Then, students read the first 10 pages of *Oedipus* and participated in a class discussion on where in the text Oedipus showed signs of pride. Students found two text examples where Oedipus was too proud and wrote them in their notes. After distributing the *Oedipus* 1 BCR prompt, the teacher read the prompt aloud. Students completed the BCR for homework and turned in the writing sample the next day for a grade. The same procedures including a warm-up, reading the text, classroom discussion, and completion of the BCR for homework were followed for the second and third prompts. Finally, the students combined all 3 *Oedipus* BCRs into one essay.

Students read *Lord of the Flies* and an accompanying study guide. As the students read over several days, they completed group work on different symbols as they appeared

in the novel. The students also completed a worksheet on symbols. As a warm-up assignment, the students selected a symbol they were interested in and found three text quotes where it appeared in the beginning, the middle, and the end. After distributing the prompt and reading the prompt aloud, the teacher and class worked together to write a thesis statement for the ECR.

All 10th-grade students included in the study completed the English 2 HSA in 2006 or 2007. Students completed the three and a half hour assessment in one school day. Classroom proctors, including classroom teachers, classroom assistants, school counselors, and administrators administered the assessment. After the Maryland State Department of Education reported the student's scores to the local district and school, the school administrator supplied the coded HSA results for each student to the investigator.

Scoring and Interscorer Reliability

To reduce possible bias in scoring caused by the quality or appearance of a student's handwriting, the investigator typed all of the writing samples exactly as written. The investigator and four school psychologists scored the writing samples. Prior to scoring the BCRs and ECRs included in the study, scorers participated in a two-hour training to ensure the consistency of the scoring and to achieve an acceptable level of reliability. The content of the training session included a review of the scoring methods, definitions, examples, guided practice in scoring five writing samples that were not included in the study, and calculation of inter-scorer agreement. End punctuation and beginning capitalization were taken into account in scoring CWS and CMIWS. The guided practice included scoring the essays together and discussing issues as they arose. Following the training, the scorers were required to score five BCRs to be included in the

study and reach a level of 90% agreement with the investigator before proceeding with scoring. Interscorer reliabilities were calculated for the scoring methods TWW, WSC, CWS, and CMIWS. Reliability coefficients were not calculated for %WSC, %CWS, or the three scoring indices measures because each of these indices is imbedded in the other scoring measures. Inter-scorer agreement was calculated by dividing the smaller score by the larger score, and multiplying by 100.

On the first attempt, the investigator and each of the scorers reached 90% or above agreement on four of the five BCRs. For BCRs that did not reach 90% agreement, the investigator and each scorer reviewed and discussed differences. After scoring the sample again, all scorers reached 90% or above agreement. To ensure that scorer drift did not occur, the investigator rescored every 10th writing sample and inter-scorer agreement was calculated. Average interscorer agreement exceeded 90% for all subsequently scored writing samples. For the double scored writing samples, the scores of the investigator were used in the final database. Total interscorer agreement between the investigator and each scorer exceeded 94%, ranging from 94.5% to 100% for all scoring methods.

In addition, each scorer timed how long it took them to score every fifth writing sample for TWW, WSC, CWS, and CMIWS. The scorer started a stopwatch before starting to score the writing sample, stopped the stopwatch, and recorded the time it took them after writing down the score. The mean amount of time needed to score a BCR or ECR was calculated for each scorer. In addition, the overall mean amount of time it took to score BCR and ECR writing samples was calculated. The average scoring time across all five scorers for the four scoring methods was calculated by adding together the means from all of the scorers and dividing by the number of scorers.

IRB Approval

The investigator obtained approval from both the University of Maryland, College Park Internal Review Board and from the local school district in which the study was completed. To protect confidentiality, a school administrator made copies of the writings and removed the student's names from the writing samples provided by the English teacher. The administrator assigned a random identification number to protect the student's identity and ensure confidentiality. Information including gender, ethnicity, special education status, and Limited English Proficient status was also provided to the investigator. The administrator kept the key with the students' names and identification numbers. The investigator did not have any method of identifying the students.

Data Analysis Procedures

The first research question asked what the alternate form reliability of written expression scoring measures and indices was for 10th-grade students for BCRs and ECRs, respectively. To address the first part of this question, for each scoring measure and index (TWW, WSC, CWS, CMIWS, %WSC, %CWS, production dependent index, and production independent index) Pearson product-moment correlations were calculated between the three *Oedipus* BCRs. The issue of writing sample length was explored to determine if scoring the first 100 words of an ECR is a reliable measure of the entire length of the text. To address this question, for each scoring measure and index (TWW, WSC, CWS, CMIWS, %WSC, %CWS, production dependent index, and production independent index), Pearson product-moment correlations between the first 100 words of "The Stone Boy" and *Lord of the Flies* ECRs and the entire length of the ECRs were calculated.

The second research question asked if there were gender differences in 10th-grade students' writing, specifically if there were differences on production dependent and production independent measures. Only students who had all five writing samples ("The Stone Boy," *Oedipus* 1, *Oedipus* 2, *Oedipus* 3, and *Lord of the Flies*) were included in this analysis. For this question, the three *Oedipus* BCRs were added together to be more equivalent to the ECRs. This question was examined using repeated measures Analysis of variance (ANOVA) with gender as a between subjects factor. Gender was the independent variable and the dependent variables included TWW, WSC, CWS, CMIWS, %WSC, %CWS, production dependent index, and production independent index.

The third research question asked what written expression scoring measures and indices were sensitive to growth in 10th-grade students' writing. To address this question, the differences between writing samples completed across the school year (Fall, Winter, and Spring) were examined. Only students who had all five writing samples were included in this analysis ("The Stone Boy," *Oedipus* 1, *Oedipus* 2, *Oedipus* 3, and *Lord of the Flies*). The sample was restricted for this question because the analysis would not be valid if a student was missing one of the writing samples. For this question, the three *Oedipus* BCRs were added together to be more equivalent to the ECRs. This question was examined using a series of repeated measures ANOVAs with time (Fall, Winter, Spring) as a within subjects factor. Due to concerns about the equivalency of the writing samples, the differences between only the two ECRs completed in the Fall and the Spring were also examined. A series of repeated measures ANOVAs with time (Fall, Spring) as a within subjects factor were run.

The fourth research question asked what written expression scoring measures and indices, were valid for predicting student performance on a state mandated high stakes assessment, the English 2 High School Assessment. This question was addressed in two different ways. First, multiple regression procedures and logistic regression procedures were run using the aggregated scores. Second, multiple regression procedures and logistic regression procedures were run using scores from the writing sample written closest in time to the administration of the English 2 HSA, the *Lord of the Flies* ECR. The multiple regression procedures were run using the student's numerical English 2 HSA score as the dependent variable and scores on scoring measures and indices as the independent variables. Logistic regression is used for examining the predictive power of one or more predictors when the outcome variable is dichotomous. The dichotomous outcome variable was passing or not passing the English 2 HSA. Scores from 396 and above were coded as "1" indicating passing. Scores ranging from 240 to 395 were coded as "0" to indicate not passing. The logistic regression prediction curve indicates the probability of being in one category or another (i.e., passing or not passing) as a function of values of the predictors.

For logistic regression models, a 2 x 2 classification table can be used to gauge the fitted model's ability to correctly predict an outcome of passing or not passing. From the parameter estimates, a predicted probability of passing or not passing was computed for each observation. Using a specific cut-point level, the predicted probabilities were converted to a predicted outcome of passing or not passing. If a predicted probability is greater than the cut-point level, the observation is predicted to be passing; otherwise, it is predicted to be not passing. The classification table shows the number of correctly

predicted outcomes and the number of incorrectly predicted outcomes. These numbers yield a percent of total observation in which the outcome of passing or not passing was correctly predicted.

Results

The research questions in this study were investigated using a series of correlational, ANOVA, and multiple regression and logistic regression analyses. A critical value equal to or less than .05 ($p < .05$) was set as the criterion for significance for all analyses. Due to restrictions placed on the study by the local school district in which data collection took place, ethnicity was not used as a variable for analysis.

Scoring Time

On average, it took almost seven minutes to score a BCR using TWW, WSC, CWS, and CMIWS with a standard deviation of over three minutes. To score an ECR using TWW, WSC, CWS, and CMIWS it took over 16 minutes with a standard deviation of almost six minutes. Appendix I presents the average scoring times for BCRs and ECRs by scorer and presents statistics (i.e. number, mean, standard deviation, minimum and maximum) on time for scoring by writing sample.

Preliminary Analyses

To test for differences among participants in Year 1 and Year 2 of the study, a series of three ANOVAs were conducted for each writing sample. The three Oedipus writing samples were added together to be equal in length to an ECR. The independent variable was year and the dependent variables included TWW, WSC, CWS, CMIWS, %WSC, %CWS, production dependent index, and production independent index

($p < .05$). For "The Stone Boy," significant differences were obtained between Year 1 and Year 2 for TWW, WSC, CWS, CMIWS, and production dependent index. For *Oedipus*, no significant differences were obtained between Year 1 and Year 2. For *Lord of the Flies*, significant differences were obtained between Year 1 and Year 2 for TWW, WSC, CWS, CMIWS, and production dependent index ($p < .001$). On "The Stone Boy" and *Lord of the Flies* writing samples, students in Year 2 of the study wrote more words, spelled more words correctly, and produced more correct minus incorrect writing sequences. Appendix J provides the three ANOVAs by year for scoring measures. Appendix K provides summary statistics by writing sample.

To investigate the interrelationships among the eight scoring measures and indices, correlational analyses were conducted by writing sample (see Appendix L). Investigation of the tables revealed that intercorrelations among the three production dependent measures (TWW, WSC, CWS) were positive and strong for all writing samples. Intercorrelations among the two production independent measures (%WSC, %CWS) were weak to moderate for all writing samples. Number of correct writing sequences correlated strongly with production dependent measures. With the exception of "The Stone Boy" writing sample, correlations between production dependent measures and production dependent measures demonstrated a negative to weak correlation.

Research Question 1: What is the reliability of written expression scoring measures indices between three BCRs?

a. What is the alternate form reliability of written expression scoring measures and indices between three BCRs?

The alternate form reliability of scoring indices was determined by examining Pearson product-moment correlations between *Oedipus* 1, *Oedipus* 2, and *Oedipus* 3 writing samples. Correlations between scoring measures and indices on the three *Oedipus* writing samples are reported in Table 5. Evaluation of the table revealed positive weak to moderate correlations for all measures and indices. The strongest correlations were found between writing samples *Oedipus* 2 and *Oedipus* 3 for TWW, WSC, CWS, CMIWS, and production dependent index (r = .67-.70, p < .001).

b. Does the reliability of written expression scoring measures differ depending on the length of the text scored?

The issue of writing sample length was explored through correlations between scoring measures and indices obtained from scoring the entire length of the ECR writing samples and scoring measures obtained from scoring the first 100 words of the ECR writing samples. Appendix M provides summary statistics for the scoring measures for the first 100 words of "The Stone Boy" and *Lord of the Flies*. Correlations between scoring measures on "The Stone Boy" and *Lord of the Flies* ECRs are reported in Table 6. For "The Stone Boy" ECR, the strongest correlations were found on %WSC, %CWS, and the production independent index (r = .87-.91, p < .001). For the *Lord of the Flies* ECR, moderate to strong correlations were found on %CWS and production independent index (r = .79, .84, p < .001). For both ECRs, the weakest correlations were obtained on production dependent measures, WSC and CWS. Low to moderate correlations were found on CMIWS for both ECRs (r = .44, .55, p < .001).

Table 5

Intercorrelations Between Oedipus Samples for Scoring Measures[a]

	Total words written			Words spelled correctly			Correct writing sequences			Correct minus incorrect writing sequences		
Sample	1	2	3	1	2	3	1	2	3	1	2	3
1. *Oedipus* 1	--	.49*	.57*	--	.51*	.58*	--	.44*	.51*	--	.39*	.48*
2. *Oedipus* 2		--	.70*		--	.69*		--	.70*		--	.67*
3. *Oedipus* 3			--			--			--			--

	% words spelled correctly			% correct writing sequences			Production dependent index[b]			Producton independent index[c]		
Sample	1	2	3	1	2	3	1	2	3	1	2	3
1. *Oedipus* 1	--	.46*	.52*	--	.40*	.59*	--	.48*	.55*	--	.44*	.61*
2. *Oedipus* 2		--	.34*		--	.46*		--	.70*		--	.47*
3. *Oedipus* 3			--			--			--			--

[a] $n = 77$. [b] Sum of total words written, words spelled correctly, and correct writing sequences. [c] Sum of % words spelled correctly and % correct writing sequences.

*$p < .01$.

Table 6

Intercorrelations Between Entire Length and First 100 Words of ECRs for Scoring Measures

	ECR	
Index	"The Stone Boy"	*Lord of the Flies*
	(*n* = 86)	(*n* = 80)
Words spelled correctly	.15	.07
Correct writing sequences	.31*	.22
Correct minus incorrect writing sequences	.55*	.44*
% words spelled correctly	.87*	.42*
% correct writing sequences	.90*	.84*
Production dependent index[a]	.17	.09
Production independent index[b]	.91*	.79*

Note. ECR = Extended Constructed Response.

[a]Sum of total words written, words spelled correctly, and correct writing sequences. [b]Sum of % words spelled correctly and % correct writing sequences.

*$p < .01$.

Research Question 2: Are there significant gender differences in 10th-grade students' writing?

a. Are there significant gender differences on production dependent measures?

b. Are there significant gender differences on production independent measures?

Appendix N provides summary statistics for the scoring measures by gender and writing sample.

Results of the repeated measures ANOVAs with gender as a between subjects factor are presented in Table 7. Results revealed significant gender differences on all

three production dependent measures, TWW, WSC, CWS, and the production dependent index. A significant gender difference was also found on CMIWS (accurate production index). Females wrote significantly more words, spelled more words correctly, produced more correct writing sequences and correct minus incorrect writing sequences. No significant gender differences were found on the production independent measures.

Research Question 3: What written expression scoring measures or indices are sensitive to growth in 10th-grade students' writing?

Results of the repeated measures ANOVAs with time (Fall, Winter, Spring) as the within subjects factor are presented in Table 7. A significant difference across time was found for the scoring measures TWW, WSC, CWS, CMIWS, %CWS, and production dependent index. No significant difference was found across time for %WSC and the production independent index.

Results of the repeated measures ANOVAs with time (Fall, Spring) as the within subjects factor are presented in Table 8. A significant difference across time was found for CWS, CMIWS, and %CWS.

Research Question 4: What written expression scoring measures or indices are valid for predicting 10th-grade student performance on a state mandated high stakes assessment?

Pearson product-moment correlation analyses between scoring measures and indices and English 2 HSA scores were conducted. Table 9 presents the correlations of writing measures and HSA scores. The aggregated data is defined as the average of each scoring measure of the two ECRs and combined BCRs ("The Stone Boy," combined *Oedipus*, and *Lord of the Flies*). A review of the table finds that the *Lord of the Flies* writing sample correlated the most strongly with HSA scores. While significant positive

Table 7

Analysis of Variance for Scoring Measures Using All Writing Samples

		F							
Source	*df*	TWW	WSC	CWS	CMIWS	%WSC	%CWS	PD[a]	PI[b]
				Between subjects					
Gender (G)	1	10.22**	9.78**	12.21**	11.90**	0.08	2.32	10.89**	1.42
Error	47	(35989)	(34468)	(30816)	(27315)	(22.06)	(153.08)	(297779)	(216.04)
				Within subjects					
Time (T)	2	3.94*	3.85*	6.91**	7.63**	0.46	4.47*	4.89**	1.41
T x G	2	0.10	0.01	0.04	0.09	0.64	1.24	0.04	1.06
Error	94	(6982)	(7159)	(6381)	(6137)	(18.33)	(22.02)	(59568)	(49.86)

Note. Values enclosed in parentheses represent mean square errors. TWW = total words written. WSC = words spelled correctly. CWS = correct writing sequences. CMIWS = correct minus incorrect writing sequences. %WSC = % words spelled correctly. %CWS = % correct writing sequences. PD = production dependent index. PI = production independent index.

[a]Sum of total words written, words spelled correctly, and correct writing sequences. [b]Sum of % words spelled correctly and % correct writing sequences.

*$p < .05$. **$p < .01$.

Table 8

Analysis of Variance for Scoring Measures Using the First and Last ECRs

		F							
Source	*df*	TWW	WSC	CWS	CMIWS	%WSC	%CWS	PD[a]	PI[b]
				Between subjects					
Gender (G)	1	5.84*	5.43*	5.42*	3.65	0.25	0.07	5.69*	0.00
Error	67	(28453)	(26833)	(25073)	(24147)	(21.50)	(147.20)	(235640)	(199.51)
				Within subjects					
Time (T)	1	1.53	0.95	5.08*	8.57**	0.16	9.43**	2.19	3.41
T x G	1	0.03	0.01	0.00	0.02	0.45	1.38	0.00	1.40
Error	67	(6395)	(6756)	(4946)	(4386)	(19.03)	(21.12)	(51960)	(56.95)

Note. The fiirst ECR was "The Stone Boy" and last ECR was *Lord of the Flies*. Values enclosed in parentheses represent mean square errors. TWW = total words written. WSC = words spelled correctly. CWS = correct writing sequences. CMIWS = correct minus incorrect writing sequences. %WSC = % words spelled correctly. %CWS = % correct writing sequences. PD = production dependent index. PI = production independent index.

[a]Sum of total words written, words spelled correctly, and correct writing sequences. [b]Sum of % words spelled correctly and % correct writing sequences.

*$p < .05$. **$p < .01$.

Table 9

Intercorrelations Between HSA Scores and Scoring Measures

Measure	"The Stone Boy"	*Oedipus*[a]	*Lord of the Flies*	Aggregates[b]
Total words written	.19	.10	.37**	.23*
Words spelled correctly	.20	.09	.33**	.22*
Correct writing sequences	.32**	.14	.42**	.32**
Correct minus incorrect writing sequences	.42**	.19	.44**	.40**
% words spelled correctly	.09	-.04	-.13	-.09
% correct writing sequences	.42**	.31*	.30*	.39**
Production dependent index[c]	.24*	.11	.38**	.26*
Production independent index[d]	.39**	.26*	.15	.32**

[a]Combined scores on *Oedipus* 1, *Oedipus* 2, and *Oedipus* 3. [b]Averages of writing scores on "The Stone Boy", *Oedipus*, and *Lord of the Flies*. [c]Sum of total words written, words spelled correctly, and correct writing sequences. [d]Sum of % words spelled correctly and % correct writing sequences.

*$p < .05$. **$p < .01$.

correlations were obtained on TWW, WSC, CWS, CMIWS, %CWS, and production dependent index, the correlations were only in the weak to moderate range. The combined *Oedipus* writing samples produced the weakest correlations with only two of the eight measures, %CWS and production independent index, producing weak but significant correlations. “The Stone Boy” writing sample correlated significantly with the HSA scores on five out of the eight scoring measures, CWS, CMIWS (accurate production index), %CWS, production dependent index, and production independent index, but again these correlations were only in the weak to moderate range. For the aggregated data, positive weak to moderate correlations were found for TWW, WSC, CWS, CMIWS (accurate production index), %CWS, production dependent index, and production independent index. Across “The Stone Boy,” *Lord of the Flies*, and aggregated data, CMIWS was moderately correlated with HSA scores.

Based on results from the correlation analyses between scoring measures and English 2 HSA scores, specific scoring measures and indices were chosen for inclusion in multiple regression and logistic regression models. Multiple regression and logistic regression analyses were run on the aggregated data and on scoring measures obtained from the *Lord of the Flies* writing sample completed in March.

Table 10 presents the results of the multiple regression analyses for the aggregated data. In the first multiple regression model, scores on the significant scoring measures TWW, WSC, CWS, and CMIWS were entered as predictors and score on the English 2 HSA was the dependent variable. The four predictors accounted for 25% of the variance in English 2 HSA scores. CMIWS ($\beta = .80$, $p < .05$) and TWW ($\beta = .98$, $p < .05$) were significant predictors. In the second multiple regression model, aggregated scores on production dependent, production independent, and accurate production (CMIWS) indices were entered as predictors and score on the English 2 HSA was the dependent variable. The three indices accounted for 21% of the variance in English 2 HSA score. Accurate production index (CMIWS) was a significant predictor ($\beta = .28$, $p < .05$). In the third multiple regression model, aggregated scores on CMIWS was entered as the predictor and score on the English 2 HSA was the dependent variable. In this model, CMIWS accounted for 16% of the variance in HSA scores and was a significant predictor ($\beta = .10$, $p < .01$).

Table 11 presents the results of the logistic regression analyses for the aggregated data. Scores on the English 2 HSA were transformed to be dichotomous with passing being coded as one and not passing coded as zero. In the first logistic regression model,

Table 10

Summary of Multiple Regression Analyses for Measures Predicting HSA Score Using Aggregated Data

Variable	*B*	*SE*
	Model 1	
Intercept	377.00**	9.11
Total words written	0.98*	0.45
Words spelled correctly	-0.41	0.22
Correct writing sequences	-1.24	0.76
Correct minus incorrect writing sequences	0.80*	0.38
	Model 2	
Intercept	498.66**	121.94
Correct minus incorrect writing sequences	0.28*	0.12
Production dependent index[a]	-0.05	0.03
Production independent index[b]	-0.67	0.67
	Model 3	
Intercept	367.28**	7.47
Corrent minus incorrect writing sequences	0.10**	0.03

Note. $N = 81$. $R^2 = 0.25$ for Model 1. $R^2 = 0.21$ for Model 2. $R^2 = 0.16$ for Model 3.

[a]Sum of total words written, words spelled correctly, and correct writing sequences.

[b]Sum of % words spelled correctly and % correct writing sequences.

*$p < .05$. **$p < .01$.

Table 11

Summary of Logistic Regression Analyses for Measures Predicting Passing the HSA Using Aggregated Data

Variable	*B*	*SE*
	Model 1	
Intercept	-2.30*	1.00
Total words written	0.26**	0.09
Words spelled correctly	-0.24**	0.09
Correct writing sequences	-0.06	0.07
Correct minus incorrect writing sequences	0.05	0.04
	Model 2	
Intercept	21.02	14.34
Correct minus incorrect writing sequences	0.03*	0.01
Production dependent index[a]	-0.01	0.00
Production independent index[b]	-0.13	0.08
	Model 3	
Intercept	-1.84**	0.71
Corrent minus incorrect writing sequences	0.01**	0.00

Note. $N = 81$. $R^2 = 0.33$ for Model 1. $R^2 = 0.20$ for Model 2. $R^2 = 0.15$ for Model 3.

[a]Sum of total words written, words spelled correctly, and correct writing sequences.

[b]Sum of % words spelled correctly and % correct writing sequences.

*$p < .05$. **$p < .01$.

aggregated scores on TWW, WSC, CWS, and CMIWS were entered as predictors. Results of the logistic regression revealed that these four predictors accounted for 32.5% of the variance in the dichotomous variable. TWW ($\beta = .26$, $p < .01$) and WSC ($\beta = -0.24$, $p < .01$) were significant predictors. Using a cut-point level of 47%, the percentage of students that did not pass the HSA, only 60.5% of actual student outcomes on the HSA were correctly predicted by this model. In the second logistic regression model, aggregated scores on production dependent, production independent, and accurate production (CMIWS) indices were entered as predictors. Results of the logistic regression revealed that these three indices accounted for 20% of the variance in the dichotomous variable. CMIWS was a significant predictor ($\beta = .03$, $p < .05$). Using a cut-point level of 47%, the percentage of students that did not pass the HSA, only 59.3% of actual student outcomes on the HSA were correctly predicted by this model. In the third logistic regression model, aggregated scores on CMIWS was entered as the predictor. Results of the logistic regression revealed that CMIWS accounted for 15% of the variance in the dichotomous variable and was a significant predictor ($\beta = .10$, $p < .01$). Using a cut-point level of 47%, the percentage of students that did not pass the HSA, only 61.7% of actual student outcomes on the HSA were correctly predicted by this model.

Table 12 presents the results of the multiple regression analyses for scores obtained from the *Lord of the Flies* (March) writing sample. In the first multiple regression model, scores on the significant scoring measures TWW, WSC, CWS, and CMIWS were entered as predictors and score on the English 2 HSA was the dependent variable. The four predictors accounted for 22% of the variance in English 2 HSA score,

Table 12

Summary of Multiple Regression Analyses for Measures Predicting HSA Score Using Lord of the Flies

Variable	*B*	*SE*
	Model 1	
Intercept	368.24**	9.76
Total words written	0.50	0.53
Words spelled correctly	-0.11	0.09
Correct writing sequences	-0.75	0.97
Correct minus incorrect writing sequences	0.48	0.48
	Model 2	
Intercept	420.03**	71.31
Correct minus incorrect writing sequences	0.15	0.08
Production dependent index[a]	-0.01	0.02
Production independent index[b]	-0.29	0.39
	Model 3	
Intercept	366.33**	7.49
Corrent minus incorrect writing sequences	0.10**	0.03

Note. $N = 68$. $R^2 = 0.22$ for Model 1. $R^2 = 0.20$ for Model 2. $R^2 = 0.19$ for Model 3.

[a]Sum of total words written, words spelled correctly, and correct writing sequences.

[b]Sum of % words spelled correctly and % correct writing sequences.

*$p < .05$. **$p < .01$.

but none of the measures was a significant predictor. In the second multiple regression model, scores on production dependent, production independent, and accurate production (CMIWS) indices were entered as predictors and score on the English 2 HSA was the dependent variable. The three indices accounted for 20% of the variance in English 2 HSA score, but none of the indices was a significant predictor. In the third multiple regression model, CMIWS was entered as the predictor and score on the English 2 HSA was the dependent variable. In this model, CMIWS accounted for 16% of the variance in HSA score and was a significant predictor ($\beta = .10$, $p < .05$).

Table 13 presents the results of the logistic regression analyses for scores obtained from the *Lord of the Flies* (March) writing sample. Scores on the English 2 HSA were transformed to be dichotomous with passing being coded as one and not passing coded as zero. In the first logistic regression model, TWW, WSC, CWS, and CMIWS were entered as predictors. Results of the logistic regression revealed that these four predictors accounted for 29% of the variance in the dichotomous variable, but none of the measures was a significant predictor. Using a cut-point level of 47%, the percentage of students that did not pass the HSA, only 61.8% of actual student outcomes on the HSA were correctly predicted by this model. In the second logistic regression model, production dependent, production independent, and accurate production (CMIWS) indices were entered as predictors. Results of the logistic regression revealed that these three predictors accounted for 23% of the variance in the dichotomous variable and none of the indices was a significant predictor. Using a cut-point level of 47%, the percentage of students that did not pass the HSA, only 61.8% of actual student outcomes on the HSA were correctly predicted by this model. In the third logistic regression model, CMIWS

Table 13

Summary of Logistic Regression Analyses for Measures Predicting Passing the HSA Using Lord of the Flies

Variable	*B*	*SE*
	Model 1	
Intercept	-2.47*	1.00
Total words written	0.09	0.08
Words spelled correctly	-0.11	0.06
Correct writing sequences	0.02	0.09
Correct minus incorrect writing sequences	0.00	0.04
	Model 2	
Intercept	14.69	14.58
Correct minus incorrect writing sequences	0.02	0.01
Production dependent index[a]	0.00	0.00
Production independent index[b]	-0.10	0.08
	Model 3	
Intercept	-1.84*	0.73
Corrent minus incorrect writing sequences	0.01**	0.00

Note. $N = 68$. $R^2 = 0.29$ for Model 1. $R^2 = 0.23$ for Model 2. $R^2 = 0.18$ for Model 3.

[a]Sum of total words written, words spelled correctly, and correct writing sequences.

[b]Sum of % words spelled correctly and % correct writing sequences.

*$p < .05$. **$p < .01$.

was entered as the predictor. Results of the logistic regression revealed that CMIWS accounted for 18% of the variance in the dichotomous variable and was a significant predictor ($\beta = .01$, $p < .01$). Using a cut-point level of 47%, the percentage of students that did not pass the HSA, only 67.6% of actual student outcomes on the HSA were correctly predicted by this model.

Discussion

The purpose of this study was to examine the use of authentic writing material collected by a classroom teacher to monitor student progress in writing and predict student performance on a state mandated high stakes assessment. The writing samples collected were based on the curriculum and written in the format required for an end of year high stakes assessment. The writing prompts were developed by classroom teachers. While the writing samples analyzed in this study were not collected using traditional CBM timed probes, the purpose of the writing was to monitor student progress over time. Further, CBM scoring measures and indices for the assessment of written expression were used to score the writing samples to determine if they would be appropriate for monitoring progress over time in teacher collected writing samples. In conducting this study, written language research with high school students was extended in the areas of reliability, validity, gender differences, and sensitivity to growth. This study was the first to use the Maryland High School Assessment as the criterion variable and to study the amount of time needed to score high school writing samples and CMIWS.

While not included as a research question, this study did investigate the total amount of time it took to score BCRs and ECRs for the scoring measures TWW ,WSC, CWS, and CMIWS. While the high interscorer reliabilities suggest that the scoring

measures were easy to use, they were also time consuming. A 10th grade English teacher may have more than 100 students across several class periods. The average scoring time of a BCR was over seven minutes with a large standard deviation of over three minutes. A review of the scoring times revealed that one BCR took over 18 minutes to score. The average scoring time of an ECR was over 16 minutes with a large standard deviation of over six minutes. While the shortest scoring time was only six minutes, the longest scoring time was almost 30 minutes. Given the low utility of the scoring measures for 10th-grade students, it would be impractical for a teacher to spend such a long amount of time scoring student writings. Given previous findings that more complex scoring measures may need to be used for high school level students it was important to look at the amount of time these measures would take to score.

Analyses conducted to determine if there was a significant difference in the writing scores of students between Year 1 and Year 2 of the study did find differences on scoring measures applied to the ECRs. Students in Year 2 of the study wrote more words, spelled more words correctly, and produced more correct minus incorrect writing sequences than students in Year 1. These differences were also borne out in the passing rates on the English 2 HSA. Significantly more students passed the assessment in Year 2 in the sample, school, district, and across the state. This large increase could be attributed to several factors. The English 2 HSA is supposed to be equivalent across years, but may have been easier for students. The HSA has been in use for several years and educators have had more time to prepare students for the assessment and students to practice taking the assessment. Students may also be more aware of the high stakes consequences

attached to the assessment. It is also possible that students in Year 2 of the study were just more skilled.

The eight measures were intercorrelated to explore the relationships between the measures and indices. Similar to previous research findings, the fluency measures, TWW, WSC, and CWS, were strongly correlated with each other and the production dependent index (Tindal & Parker, 1989, Watkinson & Lee, 1992). In addition, CMIWS was strongly correlated with the production dependent index adding to previous findings that CMIWS is tapping aspects of fluency. The high intercorrelations may be an artifact of their inherent mathematical dependencies (i.e., calculation of CWS includes both TWW and WSC) (Tindal & Parker, 1989). Similar to the previous research findings of Tindal and Parker, 1989, the correlations between production dependent measures and their production independent counterparts were only low to moderate (CWS and %CSW, *r*s = .08-.24; WSC and %WSC, *r* = .23-.42).

However, the accuracy measures were only weakly to moderately correlated with each other. This is contradictory to previous findings that the measures %WSC and %CWS are production independent or more closely related to accuracy (Tindal & Parker, 1989). The intercorrelations were only in the weak to moderate range between CMIWS (accurate production index) and the production independent index.

The first research question addressed the reliability of written language scoring measures and indices. In the first part of the research question the alternate form reliabilities of the scoring measures and indices was studied by examining the intercorrelations between three BCRs written over three consecutive weeks. Amongst all three writing samples, the *Oedipus* 2 and *Oedipus* 3 samples produced the strongest

correlations. All of the measures and indices were positive and significant ($p < .01$), but were only in the weak to moderate range. The strongest alternate form reliabilities were found on TWW, WSC, and the production dependent index. Again, while significant, the correlations were only in the weak to moderate range and were significantly lower than alternate form reliabilities obtained by Espin et al. (2000).

The second part of the research question addressed the reliability of scoring measures and indices and text length. When the text length of the ECRs was limited to the first 100 words and compared to scores on the entire text, positive strong correlations were found on both ECRS between %CWS and the production independent index. The results suggest that these production independent measures are good measures of the writing accuracy of the entire length of the writing sample.

The second research question addressed gender differences on production dependent and production independent measures. Similar to results obtained by Jewell and Malecki with a sample of elementary level students, this study found that boys and girls differed on TWW, WSC, and CWS, and the production index. However, a significant difference was also found on CMIWS (accurate production index) that was not found by Jewell and Malecki. On all writing samples, girls wrote more words, spelled more words correctly, produced more correct writing sequences, and produced more correct minus incorrect writing sequences. These results lend support to findings that while boys and girls differ in writing production, they do not differ in writing accuracy.

The third research question addressed what scoring measures and indices are sensitive to growth. While a significant difference for time was found between the fall, winter, and spring writing samples, there was not a consistent increase in these measures

across the school year. A review of the statistics for scoring methods tables found an increase on measures from fall to winter, but then a decrease from winter to spring. This could be related to the fact that the three *Oedipus* BCRs were combined together to be more equivalent in length to the two ECRs. The decision to just look at growth between the two ECRs completed the fall and the spring found an increase in CWS, CMIWS, and %CWS. While these measures may be sensitive to growth, the growth was not large enough to provide information useful for monitoring progress. For example, an increase in 25 CMIWS from the fall to the spring means an increase in approximately 3 CMIWS per month. However, the increase in growth is interesting in light of the fact that the teacher was not following an explicit program of writing instruction. More growth might have resulted if the teacher had delivered writing instruction. In general, very little writing instruction takes place at the high school level (Espin et al., 2005; Parker et al., 1991b). The small changes in growth are not useful for instructional decision making and planning (Espin et al., 2005; Parker et al., 1991b). These small changes are not instructionally useful enough to merit the excessive amount of time needed to scores these measures.

The fourth research question addressed the validity of scoring measures or indices for predicting student performance on a state mandated high stakes assessment. The magnitude of correlations between the scoring measures and indices were weak and lower than validity coefficients found in previous studies. One possible explanation for the difference is the age of the students in this study. The students in this study were in 10th grade whereas students in previous studies have been in middle school. It is possible that as students become older and more proficient in written expression, the validity of

the scoring measures and indices decreases. This hypothesis is supported by Parker et al. (1991b) and Weissenburger and Espin (2005) who found decreases in correlations between scoring measures and criterion measures with increases in student age.

Results of multiple regression and logistic regression analyses failed to provide a model that could accurately predict student outcomes on the English 2 HSA. If a model was found that could accurately predict student outcomes, then a table of probable success on the HSA could have been created (Espin et al., 2006). The accuracy of such a table depends on the strength of the relationship between the predictor and the assessment outcome (Espin et al., 2006). The finding that these measures and indices were not valid predictors of performance precluded taking the next step of providing a table of probable of success that could be used to identify students at-risk for not passing the assessment.

Implications for Practice

One of the interests of this study was to provide ecologically valid information on classroom practices and to inform the instruction of teachers. In terms of implications for practice, the results suggest that scoring measures and indices used in curriculum-based measurement studies may not be valid for use with secondary students. While previous research completed by Espin et al. and Jewell and Malecki (2005) provided evidence that more complex scoring measures such as CMIWS are technically sound for older students, the computation of these scores takes an impractical amount of time when the entire length of an ECR needs to be scored.

Findings that boys and girls differed significantly on production dependent measures have implications for educators. Given these differences, separate norms on fluency measures for boys and girls may be needed (Jewell & Malecki, 2005; Malecki

and Jewell, 2003). Use of the same fluency norms for boys and girls could lead to the over identification of boys for writing difficulties, especially in the early grades where simpler fluency measures have been found to be technically adequate.

Implications for Research

In this time of increased accountability and calls for research-based methods to monitor student progress and inform instruction, there is a clear need for further research in written expression. Previous research findings on the technical adequacy of written expression scoring measures and indices have laid a foundation for future research on progress measures. With this research, students at-risk for failing to meet state assessment standards could be identified, interventions implemented, and student progress monitored (McMaster & Espin, 2007).

The technical adequacy of progress measures in written expression for high school students needs further investigation with both a larger sample of students and across grade levels. Studies across grade levels are important for more long-term progress monitoring and goal setting toward meeting state standards.

Future studies investigating the criterion validity of other measures of written expression may provide valuable. Early research conducted by Deno and colleagues found stronger criterion validity coefficients than more recent research at the elementary and secondary levels (McMaster & Espin, 2007). The difference may be caused by criterion measures that do not directly assess written expression or due to the previously discussed difficulties of measuring written expression. McMaster and Espin pointed out that although validity coefficients for writing measures have been lower than those seen for measures in other areas (reading and mathematics) coefficients in many of the studies

are similar to or better than those seen for other commonly used measures of written expression including the TOWL and Woodcock Johnson Tests of Achievement.

This study, like previous studies, used writing samples collected across a long period of time. In order to measure changes in performance and progress, an important goal would to develop measures that could be used to monitor progress on a more frequent basis to facilitate instructional decisions. At both the elementary and secondary levels there is a need for more research on measuring progress over time. Findings of variations in scores obtained from different writing prompts has implications for future research addressing the use of measures for monitoring progress (McMaster & Espin, 2007). Variability in interest and background knowledge may impact the quantity and quality of student's written responses. This variability would lead to "substantial bounce" from one data point to the next, negating the utility of the measures to monitor progress over time (Parker et al., 1991b).

Lastly, the school district where the data collection took place would not allow analysis of the data in regards to ethnicity. Additional research on the impact of student ethnicity, gender, limited English proficiency status, and educational disability is warranted.

Limitations

Several potential limitations of the study require consideration.

A first limitation of the study was the use of only one criterion variable, a high stakes end of year assessment. It is difficult to identify a criterion measure that will provide a measure of written expression skills. The English 2 HSA scores are a combination of student selected response, brief constructed response, and extended

constructed response scores. The English 2 HSA may not be a good measure of written expression skills.

A second limitation of the study was the sample. The writing samples in the study were obtained from two different academic years, and the sample was limited in size. While the small sample size may be considered a limitation of the study, if significant results cannot be obtained using a small sample, then these measures will not be useful for making instructional decisions on an individual student level.

A third limitation of the study was the writing samples. A review of cumulative writing folders found that classroom teachers had assigned different writing topics. In order to compare student writings on the same topics, it was determined that only writings on the same topic would be utilized in this study. This led to the inclusion of only students from one classroom teacher. The classroom teacher with the highest number of writing samples was chosen for the study. Many students had assignments missing from their folders indicating that they might not have completed the assignment or just did not put the assignment in their writing folder. The writing assignments were written on book topics and a poor written response may reflect that the student did not read the book or possessed poor reading skills. The reading skills of the students were not assessed as part of this study. Another limitation was that the writing samples were not written only during class time and students worked on the writing assignments at home.

A fourth limitation was the classroom teacher. The selection of students enrolled in the teacher's classes may not have been random. The lack of relationship may be attributable due to her particular teaching style, selection of writing assignments, or the

nature of these assignments. The classroom teacher provided information on the presentation of the writing prompts and explanation of the writing assignments.

A final limitation of the study was the potential threat to external validity and limited generalization to other populations due to the fact that this study was conducted in a suburban school district.

Conclusion

Results from this study failed to provide evidence of the technical adequacy of scoring measures for high school students. It may be that more "traditional" scoring measures and indices shown to be technically adequate for elementary and middle school students are not technically adequate for high school students. Additional research is needed to examine the technical adequacy of scoring measures at the high school level and with other criterion measures of written expression. Enabling teachers to efficiently and effectively monitor the writing skills of their students remains a critical but elusive area for study.

Appendix A

Table of Written Expression Studies

Study	Grade(s)	Type of Prompt	Time (min)	Scoring Methods	Criterion Validity	Results & Limitations
			Elementary School Studies			
Deno, Marston, & Mirkin (1982)	3-6	Story prompt Picture stimulus	1-5	TWW WSC Large words Mature words Mean T-Unit CLS	TOWL SAT DSS	Criterion validity with TOWL: TWW .41-.82; WSC .45-.88 Large words .29-.72 Mature words .41-.79 Mean T-unit .03-.33 Criterion validity with TOWL & SAT: TWW .57-.81; WSC .60-.80 Mature words .60-.88 Large words .50-.75 T-unit length .02-.58 Criterion validity with DSS: TWW .65-.88; WSC .67-.84 CLS .64-.86; Mature words .54-.74 Large words .23-.35 T-unit length .29
Videen, Deno, & Marston (1982)	3-6	Story Topic sentence	5	CWS	DSS TOWL Holistic rating Mean T-unit Poteet checklist	Interscorer reliability .90 Criterion validity correlations coefficients: DSS .49 TOWL .69 Holistic rating .85 Mean T-unit .18 Poteet checklist -.03 to .20

(Appendix A continued)

Study	Grade(s)	Type of Prompt	Time (min)	Scoring Methods	Criterion Validity	Results & Limitations
Deno, Marston, Mirkin, Lowry, Sindelar, & Jenkins (1982)	1-6	Story	3	TWW WSC CLS		Interscorer reliability: .96-.99 Growth stability correlations: TWW .27-.72 WSC .20-.78 CLS .36-.86
Parker, Tindal, & Hasbrouck (1991a, Study 1)	2-5	Story	6	TWW WSC CWS %WSC %CWS	Teachers' holistic ratings	Weak to moderate correlations with holistic ratings. %CWS most viable screening tool. Other scoring measures lacked sensitivity to student differences. Limitation: Students stopped writing before 6 minutes were up.
Tindal & Parker (1991)	3-5	Story	3-10	TWW WSC CWS	Analytic scoring system	Interscorer reliability: .92-.99 Criterion validity correlation coefficients: -.02 to .63 Correlation with holistic judgments: $r = .85$ Correlations between TWW, WSC, & CWS and SAT: $r = .18$-$.41$ Significant improvement over time. Differences between groups.

(Appendix A continued)

Study	Grade(s)	Type of Prompt	Time (min)	Scoring Methods	Criterion Validity	Results & Limitations
Gansle et al. (2002)	3, 4	Story	3	TWW, WSC CWS No. of nouns, verbs, adjectives Long words Total punctuation Correct punctuation Capitalization Complete sentences Sentence fragments Simple Sentences	Teacher rating ITBS LEAP	Alternate form reliability: .006-.62 Criterion validity correlation coefficients: Teacher ratings -.14 to .37 ITBS -.24 to .36 LEAP -.12 to .33
Gansle et al. (2004)	3, 4	Story	3	TWW, CWS Punctuation Marks, Correct Punctuation, Words in Complete Sentences, Simple sentences	Woodcock-Johnson Revised Writing Sample Subtest	Criterion validity correlation coefficients: TWW .23; Punctuation marks .42; Correct punctuation .34; Words in complete sentences .35; CWS .36; Simple sentences -.05

(Appendix A continued)

Study	Grade(s)	Type of Prompt	Time (min)	Scoring Methods	Criterion Validity	Results & Limitations
			Middle School Studies			
Tindal & Parker (1989)	6, 7, 8	Story	3, 6	TWW, WSC CWS, LW ML/CWS %WSC %CWS %LW	Teachers' holistic ratings	Criterion validity correlation coefficients: TWW .10; WSC .24; CWS .31; LW .45; ML/CWS .59; %WSC .42; %CWS .73; %LW .75 Limitations: Limited generalization. Holistic rating of writing.
Parker, Tindal & Hasbrouck (1991a, Study 2)	6, 8, 11	Story	6	TWW WSC CWS %WSC %CWS	Holistic ratings	Interrater reliability: .87-.99 Criterion validity correlation coefficients: TWW .39-.41; WSC .43-.52; CWS .48-.56; %WSC .34-.46; %CWS .36-.42 Limitations: Students stopped writing before 6 minutes were up. Use of holistic rating as sole criterion. Reliability of holistic rating not high. Only 1 writing sample.

(Appendix A continued)

Study	Grade(s)	Type of Prompt	Time (min)	Scoring Methods	Criterion Validity	Results & Limitations
Parker, Tindal & Hasbrouck (1991b)	6-8	Story	6	TWW WSC LW CWS ML/CWS %WSC %LW	Holistic ratings Test of Written Language	Internal reliability .60-.81 Growth stability: TWW .69-.83; LW .69-.83; WSC .68-.79; CWS .49-.77; ML/CWS .26-.66; %LW .17-.76; %CWS .45-.75 Criterion validity: Holistic: TWW .39; LW .45; WSC .54; CWS .64; ML/CWS .63; %LW .60; %CWS .53 TOWL: TWW .16; LW .56; WSC .25; CWS .27; ML/CWS .18; %LW .56; %CWS .28 Limitations: Small sample size. Single writing sample at each point. Students finished before 6 minutes.
Watkinson & Lee (1992)	6, 7, 8	Story	3	TWW LW WSC CWS IWS %LW %WSC %CWS	Intercorrelations with production dependent measures Intercorrelations with production independent measures	Interscorer reliability .80-.99 Significant differences between Non-LD and LD students on CWS, IWS, and production independent measures Intercorrelations among 4 production dependent variables positive and strong (.77-.99). Intercorrelations between 3 production independent measures positive and strong (.79-.92) CWS positive moderate correlation with production-independent variables. Limitations: Only 52 students.

(Appendix A continued)

Study	Grade(s)	Type of Prompt	Time (min)	Scoring Methods	Criterion Validity	Results & Limitations
Espin, Shin, Deno, Skare, Robinson, & Benner (2000)	7, 8	Story Expository	3, 5	TWW WSC Characters Sentences CWS CMIWS	Teacher ratings District test	Alternate form reliability: strongest for TWW, WSC, CWS, CMIWS (r = .72 to r = .80) Coefficients similar across type of writing and duration. Criterion validity correlation coefficients: Low to moderate range. CMIWS most strongly related to teacher rating and performance on district test. Limitations: One teacher rated samples. Only 8th grade took district test.
Espin, De La Paz, Scierka, & Roelofs (2005)	7, 8	Expository	35	TWW CWS CMIWS	Functional Elements Holistic ratings	Correlations between TWW, CWS CMIWS and criterion were moderate to strong (rs = .58 - .90). Validity coefficients for first 50 words resulted in decreased correlations (rs = .33 - .59). TWW, CWS, CMIWS increased over time. Increases in CWS and CMIWS observed for lower performers. Longer samples needed to detect growth in higher performers. Limitations: Small sample size. Essays typed and corrected. Pre-test and post-test design for measuring change in performance.

(Appendix A continued)

Study	Grade(s)	Type of Prompt	Time (min)	Scoring Methods	Criterion Validity	Results & Limitations
				High School Study		
Espin, Scierka, Skare, & Halverson (1999)	10	Story	3	TWW WSC CWS Character Sentences ML/CWS	English GPA Teacher Ratings California Achievement Test	Low to moderate validity coefficients (r = .30-.45) 4 groups differed significantly on characters per word, sentences, ML/CWS Combination of characters, sentences, and ML/CWS provided best index of writing. Limitations: No measure of reliability. Investigator typed samples. Only one writing sample. No validity data on CAT. CAT completed in 11th grade and samples completed in 10th grade.
				Studies Across Grade Levels		
Jewell & Malecki (2003)	1-8	Story	3	TWW WSC CWS CMIWS %WSC %CWS		Interscorer reliability: .98-.99 Grade level differences on TWW, WSC, CWS, &WSC, %CWS, CMIWS Gender main effect – girls outperformed boys on all measures Intercorrelations: with older students how much they wrote was not closely related with accuracy of their writing.

(Appendix A continued)

Study	Grade(s)	Type of Prompt	Time (min)	Scoring Methods	Criterion Validity	Results & Limitations
Jewell & Malecki (2005)	2, 4, 6	Story	3	TWW WSC CWS %WSC %CWS CMIWS	THASS SAT Language Arts Grade	Girls outperformed boys on all writing fluency measures. No gender differences on production-independent or accurate production indices. Limitations: Lack of data collection integrity data. Samples only scored by one rater.
Weissenburger & Espin (2005)	4, 8, 10	Story	3, 5, 10	TWW CWS CMIWS	WKCE Language Arts Test	Alternate form reliability: TWW .55-.84; CWS .59-.84; CMIWS .61-.82 Criterion validity correlation coefficients: TWW .26-.28; CWS .39-.40; CMIWS .46-.48 Limitations: Sample 96% Caucasian Lack of complete data set. Holistic scores not available for 10th grade students.

Appendix B

Brief Constructed Response Rubric

Level 3

The response demonstrates an understanding of the complexities of the text.

- Addresses the demands of the question
- Uses expressed and implied information from the text
- Clarifies and extends understanding beyond the literal

Level 2

The response demonstrates a partial or literal understanding of the text.

- Addresses the demands of the question, although may not develop all parts equally
- Uses some expressed or implied information from the text to demonstrate understanding
- May not fully connect the support to a conclusion or assertion made about the text(s)

Level 1

The response shows evidence of a minimal understanding of the text.

- May show evidence that some meaning has been delivered from the text
- May indicate a misreading of the text or question
- May lack information or explanation to support understanding of the text in relation to the question.

Level 0

The response is completely irrelevant or incorrect, or there is no response.

(Maryland State Department of Education, 2006)

Appendix C

Extended Constructed Response Rubric

Level 4

The response is a well-developed essay that fulfills the writing purpose.

- Develops ideas using relevant and complete support and elaboration
- Uses an effective organizational structure
- Uses purposeful word choice
- Demonstrates attention to audience's understanding and interest
- Has no errors in usage or conventions that interfere with meaning

Level 3

The response is a complete essay that addresses the writing purpose.

- Develops ideas using adequate support and elaboration
- Uses an organizational structure that supports the writing. Purpose
- Uses clear word choice
- Demonstrates an awareness of audience's understanding and interest
- Has few, if any, errors in usage and conventions that interfere with meaning

Level 2

The response is incomplete or oversimplified attempt to address the writing purpose.

- Has incomplete or unclear support and elaboration
- Attempts to use an organizational structure
- Demonstrates little awareness of audience's understanding and interest
- May have errors in usage and conventions that interfere with meaning

Level 1

The response provides evidence of an attempt to address the prompt.

- Has minimal or no support or elaboration
- May be too brief to demonstrate an organizational structure
- Demonstrates little or no awareness or audience
- May have errors in usage and conventions that interfere with meaning

Level 0

The response is completely irrelevant or incorrect.

(Maryland State Department of Education, 2006)

Appendix D

Scoring BCRs and ECRs by the MSDE

According to Swanson (2007), a lengthy process is used to select, train, and hire all scorers of the writing samples. Scorers are initially hired to participate only in training sessions. If they meet the training criteria, then they are hired to score real student writing responses. Each potential scorer is interviewed twice and must have a minimum of a Bachelor's Degree and provide a writing sample. A Scoring Director is in charge of all l00 scorers hired. The Scoring Director works with a team of 10 to 12 Team Leaders who are trained as a group before training sessions for scorers begin. Each Team Leader is responsible for a group of 10 to 12 scorers. After a review of the guidelines, practice papers are read by the potential scorers and are discussed. Then, each potential scorer scores 4 sets of 20 qualifying student responses, which have been previously scored by the scoring committee. The potential scorer must achieve 80% "perfect agreement" on the qualifying scoring sets before being hired.

Scorers work independently to score real student responses, but each scorer is also part of a team (Swanson, 2007). Two raters from different teams score each response and responses are not discussed as a group. Statistical records are kept on all scorers. Scorers assign only whole number scores to the response. If scores are adjacent, within a one-point discrepancy, the average of the two scores is assigned. However, if scores are not within one point, a third "expert scorer," Team Leader or Scoring Director, scores the response and the student receives the score assigned by the expert. Validity sets, similar to training and qualifying sets, are also circulated that have already been scored by the scoring committee. The validity sets are utilized to limit "scorer drift" and provide

additional feedback to the scorer. The scoring rubric is not the only scoring tool utilized. The rubric is used in combination with model or "anchor papers" that illustrate the application of the strengths or weaknesses of the criteria to any given score point for each BCR or ECR. The scorer works to match the student response to the anchor paper and the rubric.

Appendix E

BCR and ECR Item Writing Guidelines and Sample Items

Item Writing Guidelines

The item should :

1. clearly tell students what they are to do and what is expected of them.

2. clearly tell students where they are to write their response. The amount of space provided on the answer document should be appropriate for the length of the expected response.

3. use simple but authentic vocabulary and good sentence structure. It should be clear and concise.

4. identify the information or material that students should use when preparing their response. It should focus their attention on the particular area of knowledge or to the specific aspects of the stimulus material that the students should use.

5. clearly indicate the scientific skill or process that should be demonstrated in the response.

6. provide proper cueing to direct the student's thinking and identify expectations.

Public Release HSA 2007
http://hsaexam.org/sample/english.html

Sample English 2 HSA BCR:
Reading the essay "A Sea Worry."
Carefully examine the details in the photographs provided below.
Write a response that explains which photograph better communicates ideas similar to the ideas expressed in the essay "A Sea Worry." In your response, support your conclusion with appropriate details from both the essay and photograph you choose.

Use the space on page ___ of your Answer Book for planning your response.
Then write your response on the lines on page ___.

Sample English 2 HSA ECR:
Consider the following statement by author Edward Alden Jewell:
"To paint a picture is far more important than to sell it."
Write a well-organized essay in which you agree or disagree with Jewell's statement. Support your position with specifi c examples from your studies, experiences, or observations. Be sure that your essay is fully developed, that it is logically organized, and that your choice of words clearly expresses your ideas.

Use the space on page ___ in your Answer Book for planning your essay. Then write your essay on the lines on pages ___ and ___.

(Maryland State Department of Education, 2006)

Appendix F

"The Stone Boy" Prompt

In a well thought out paper, compare the poem "Weapons," by McGinley with the short story "The Stone Boy," by Gina Berriault. What do these two works have in common? How is the theme related by both expressed?

Theme: A person's misunderstood behavior or actions can lead to their isolation from society.

I. Introduction:

- a. Short summary, titles, author
- b. thesis
- c. controls

II. Event # One:

- a. example with text quote

III. Event # Two:

- a. Example with text quote

IV. Event # Three:

- a. Example with text quote

V. Conclusion:

- a. restate main points
- b. explain how this shows a "broken heart" (relate back to the poem)

Appendix G

Oedipus Writing Prompts

Oedipus BCR 1 – Oedipus has a tragic flaw

10 sentence BCR on Oedipus' Pride…

Oedipus' tragic flaw is pride. Analyze and explain how Oedipus suffers from being too proud. Give at least two text quotes as evidence to support your claim and explain how they demonstrate his excessive pride.

Oedipus BCR 2 - Anagnorisis

Write a 10 sentence BCR explaining Oedipus' anagnorisis (his realization that he himself has caused his own downfall). Be sure to include the appropriate text quote on page 89, Oedipus – "Oh God! It has all come true. Light let this be the last time I see you. I stand revealed – born in shame, married in shame, and unnatural murderer." Explain the events that lead up to this moment and how the quote shows Oedipus' understanding that his own pride caused his own downfall.

(Mini-Outline)

Topic sentence: Oedipus becomes aware of his anagnorisis in the middle of the play.

Anagnorisis is when a character recognizes his contribution to his own downfall.

Context – Summary of what is going on in the play.

Evidence – Text quote.

Explanation – Explain quote, events that lead up to it, and how he understands it is his pride that was his downfall.

Oedipus BCR 3 - Restoration

Write a 10 sentence BCR explaining how the play shows the concept of restoration needed at the end of a tragedy. What is the outcome of Oedipus' realization on page 89? How does he react to his understanding? What consequences are there for his life and that of his children?

(Mini-Outline)

Topic sentence: Order is restored at the end of the play when Oedipus takes responsibility for his prideful actions.

Context – Summary of the end of the play.

Evidence – Text Quote "From this I am cut off, I, the most nobly raised in Thebes, cut off by my own act" (98).

Explanation – Explain the outcome of Oedipus' knowledge that he caused his fate – how does he punish himself? Explain quote and how it shows his acceptance of his fate. Explain how society is restored in the end of the play, and finally how the audience learns from Oedipus' story.

Appendix H

Lord of the Flies Prompt

Your final essay will focus on tracing a symbol as it develops through the beginning, middle, and end of the novel. You may choose any symbol that is present in the novel to complete this essay. Please keep in mind the ironic movement of the novel; things move from good to bad.

1. Choose a symbol from the novel and decide what it represents:

    Conch

    Glasses

    Clothing

    Fire

    Beast

    Island

    Spears

2. Decide whether the change is through function of the item, or physical description of the item.

3. Examine how the symbolic relationship is shown through the boys' actions on the island.

Your essay will consist of at least five paragraphs using the following organizational outline:

1. An introduction that discusses your item and what abstract concept it symbolizes through your thesis statement. Please include: the object, what it symbolizes, and how it change from the beginning to the end of the novel.

2. Three controls that include:

- A topic sentence that introduces the status of the symbol and its meaning at the part of the book.
- A specific text reference to the state of the symbol
- How the state of the symbol represents the symbolic idea on the island
- A specific text reference to how the boys' actions support your interpretation of the symbol at this point in the book.

3. A conclusion that summarizes your thesis and your main points and provides a sense of closure to your paper.

Appendix I

Statistics on Scoring Time

Table H1

Average Scoring Time by Scorer

Scorer	*M*	*SD*
	BCRs	
1	6.96	2.16
2	9.67	3.17
3	6.88	2.94
4	5.11	2.41
	ECRs	
1	14.29	4.97
2	17.33	5.99
3	11.26	2.51

Note. Scoring time is measured in minutes. BCR = Brief Construction Response; ECR = Extended Constructed Response.

Table I2

Statistics on Scoring Time by Writing Sample

Sample	*N*	*M*	*SD*	*Min*	*Max*
Oedipus 1	32	7.81	3.01	2.15	16.30
Oedipus 2	20	8.07	3.23	4.48	18.48
Oedipus 3	48	5.85	2.71	1.40	12.18
"The Stone Boy"	41	16.43	5.62	6.13	28.53
Lord of the Flies	37	16.66	5.76	8.12	29.30

Note. Scoring time is measured in minutes.

Appendix J

Analyses of Variance by Year for Scoring Measures

Source	*df*	*F*							
		TWW	WSC	CWS	CMIWS	%WSC	%CWS	PD[a]	PI[b]
The Stone Boy									
Year	1	8.02**	7.69**	7.73**	5.61*	0.53	0.39	7.97**	0.20
Error	84	(17892)	(17622)	(15762)	(15789)	(2.81)	(97.39)	(150766)	(122.00)
Oedipus									
Year	1	0.14	0.15	0.08	0.02	0.02	0.06	0.12	0.06
Error	75	(24125)	(23714)	(21655)	(18381)	(2.97)	(54.74)	(206300)	(71.64)
Lord of the Flies									
Year	1	22.21**	18.40**	17.85**	9.94**	1.90	0.16	20.23**	0.99
Error	78	(13690)	(13392)	(12705)	(13324)	(33.69)	(75.58)	(114822)	(134.52)

Note. Values enclosed in parentheses represent mean square errors. TWW = total words written. WSC = words spelled correctly. CWS = correct writing sequences. CMIWS = correct minus incorrect writing sequences. %WSC = % words spelled correctly. %CWS = % correct writing sequences. PD = production dependent index. PI = production independent index. [a]Sum of total words written, words spelled correctly, and correct writing sequences. [b]Sum of % words spelled correctly and % correct writing sequences.

*$p < .05$. **$p < .01$.

Appendix K

Summary Statistics by Year for Scoring Measures

Table K1

Summary Statistics for Scoring Measures for "The Stone Boy"

Measure	*M*	*SD*	*Min*	*Max*
Overall (*N* = 86)				
Total words written	368.3	139.2	151	737
Words spelled correctly	361.7	137.9	150	725
Correct writing sequences	318.8	130.4	117	684
Correct minus incorrect writing sequences	249.6	129.0	35	627
% words spelled correctly	98.1	1.7	92.5	100.0
% correct writing sequences	81.6	9.8	57.7	96.5
Production dependent index	1048.8	403.9	447	2118
Production independent index	179.7	11.0	153	196
Year 1 (*n* = 45)				
Total words written	329.3	135.7	151	710
Words spelled correctly	323.8	134.3	150	707
Correct writing sequences	282.9	126.1	117	684
Correct minus incorrect writing sequences	218.9	125.3	35	627
% words spelled correctly	98.2	1.8	92.5	100.0
% correct writing sequences	81.0	11.1	57.7	96.5
Production dependent index	936.0	392.2	447	2101
Production independent index	179.2	12.4	153	196
Year 2 (*n* = 41)				
Total words written	411.1	131.6	206	737
Words spelled correctly	403.3	131.1	201	725
Correct writing sequences	358.3	125.0	181	656
Correct minus incorrect writing sequences	283.2	126.0	92	569
% words spelled correctly	98.0	1.5	93.2	100.0
% correct writing sequences	82.3	8.3	61.4	94.7
Production dependent index	1172.7	384.0	588	2118
Production independent index	180.3	9.4	159	194

Note. Production dependent index is sum of total words written, words spelled correctly, and correct writing sequences. Production independent index is sum of % words spelled correctly and % correct writing sequences.

Table K2

Summary Statistics for Scoring Measures for Oedipus 1

Measure	*M*	*SD*	*Min*	*Max*
Overall (*N* = 77)				
Total words written	117.2	45.8	35	281
Words spelled correctly	115.5	45.4	34	281
Correct writing sequences	106.5	44.6	34	268
Correct minus incorrect writing sequences	88.3	44.0	11	240
% words spelled correctly	98.4	2.0	89.8	100.0
% correct writing sequences	85.0	9.6	55.1	98.7
Production dependent index	339.2	135.0	103	830
Production independent index	183.4	10.8	151	199
Year 1 (*n* = 37)				
Total words written	104.1	42.5	43	256
Words spelled correctly	102.4	42.0	42	255
Correct writing sequences	92.0	40.8	36	238
Correct minus incorrect writing sequences	73.4	40.5	11	205
% words spelled correctly	98.3	2.0	89.8	100.0
% correct writing sequences	82.8	10.8	55.1	95.7
Production dependent index	298.5	124.1	122	749
Production independent index	181.1	11.9	151	196
Year 2 (*n* = 40)				
Total words written	129.4	45.9	35	281
Words spelled correctly	127.6	45.7	34	281
Correct writing sequences	119.9	44.4	34	268
Correct minus incorrect writing sequences	102.1	43.1	29	240
% words spelled correctly	98.6	2.0	89.8	100.0
% correct writing sequences	87.0	7.9	67.0	98.7
Production dependent index	376.9	135.1	103	830
Production independent index	185.6	9.3	157	199

Note. Production dependent index is sum of total words written, words spelled correctly, and correct writing sequences. Production independent index is sum of % words spelled correctly and % correct writing sequences.

Table K3

Summary Statistics for Scoring Measures for Oedipus 2

Measure	*M*	*SD*	*Min*	*Max*
Overall (*N* = 77)				
Total words written	166.2	75.6	37	394
Words spelled correctly	163.2	74.8	35	386
Correct writing sequences	152.1	72.7	32	363
Correct minus incorrect writing sequences	123.7	66.4	16	302
% words spelled correctly	98.0	2.2	87.0	100.0
% correct writing sequences	84.2	8.7	52.7	98.1
Production dependent index	481.5	221.9	107	1130
Production independent index	182.2	9.7	153	198
Year 1 (*n* = 37)				
Total words written	167.9	63.9	39	304
Words spelled correctly	164.9	63.0	39	300
Correct writing sequences	157.4	65.4	32	363
Correct minus incorrect writing sequences	133.0	58.9	21	302
% words spelled correctly	98.2	1.7	93.0	100.0
% correct writing sequences	86.3	7.5	66.3	98.1
Production dependent index	490.2	190.8	110	967
Production independent index	184.5	8.5	162	198
Year 2 (*n* = 40)				
Total words written	164.6	85.7	37	394
Words spelled correctly	161.7	85.1	35	386
Correct writing sequences	147.2	79.4	35	350
Correct minus incorrect writing sequences	115.1	72.3	16	293
% words spelled correctly	97.8	2.6	87.0	100.0
% correct writing sequences	82.2	9.4	52.7	97.8
Production dependent index	473.5	249.4	107	1130
Production independent index	180.1	10.3	153	198

Note. Production dependent index is sum of total words written, words spelled correctly, and correct writing sequences. Production independent index is sum of % words spelled correctly and % correct writing sequences.

Table K4

Summary Statistics for Scoring Measures for Oedipus 3

Measure	*M*	*SD*	*Min*	*Max*
Overall (*N* = 77)				
Total words written	140.9	58.4	34	261
Words spelled correctly	138.4	57.6	33	258
Correct writing sequences	128.8	55.3	22	254
Correct minus incorrect writing sequences	108.0	51.9	0	239
% words spelled correctly	98.2	2.2	87.0	100.0
% correct writing sequences	85.7	9.1	50.0	97.9
Production dependent index	408.1	170.5	89	755
Production independent index	183.9	10.5	141	198
Year 1 (*n* = 37)				
Total words written	145.5	57.9	62	258
Words spelled correctly	142.9	56.6	61	247
Correct writing sequences	132.9	54.2	39	240
Correct minus incorrect writing sequences	111.2	49.8	0	207
% words spelled correctly	98.3	1.9	90.7	100.0
% correct writing sequences	85.5	9.1	50.0	97.9
Production dependent index	421.3	167.8	169	732
Production independent index	183.8	10.4	141	197
Year 2 (*n* = 40)				
Total words written	136.6	59.3	34	261
Words spelled correctly	134.3	59.1	33	258
Correct writing sequences	124.9	56.8	22	254
Correct minus incorrect writing sequences	105.1	54.3	10	239
% words spelled correctly	98.2	2.5	87.0	100.0
% correct writing sequences	85.9	9.2	64.4	97.6
Production dependent index	395.8	174.1	89	755
Production independent index	184.1	10.7	158	198

Note. Production dependent index is sum of total words written, words spelled correctly, and correct writing sequences. Production independent index is sum of % words spelled correctly and % correct writing sequences.

Table K5

Summary Statistics for Scoring Measures for Aggregated Oedipus

Measure	*M*	*SD*	*Min*	*Max*
Overall (*N* = 77)				
Total words written	424.3	154.4	106	936
Words spelled correctly	417.1	153.1	102	925
Correct writing sequences	387.4	146.3	91	854
Correct minus incorrect writing sequences	320.0	134.7	68	722
% words spelled correctly	98.2	1.7	91.5	100.0
% correct writing sequences	85.0	7.4	65.6	97.2
Production dependent index	1228.8	451.6	299	2715
Production independent index	183.2	8.4	157	197
Year 1 (*n* = 37)				
Total words written	417.5	142.5	171	729
Words spelled correctly	410.1	140.0	169	715
Correct writing sequences	382.4	136.3	139	698
Correct minus incorrect writing sequences	317.6	125.7	70	585
% words spelled correctly	98.2	1.6	91.5	100.0
% correct writing sequences	85.2	7.7	65.6	94.9
Production dependent index	1210.0	416.1	479	2085
Production independent index	183.4	8.8	157	194
Year 2 (*n* = 40)				
Total words written	430.6	166.3	106	936
Words spelled correctly	423.6	165.9	102	925
Correct writing sequences	392.0	156.5	91	854
Correct minus incorrect writing sequences	322.3	144.1	68	722
% words spelled correctly	98.2	1.8	93.2	100.0
% correct writing sequences	84.8	7.1	66.7	97.2
Production dependent index	1246.2	486.7	299	2715
Production independent index	183.0	8.2	161	197

Note. Aggregated *Oedipus* is the sum of *Oedipus* 1, *Oedipus* 2, and *Oedipus* 3. Production dependent index is sum of total words written, words spelled correctly, and correct writing sequences. Production independent index is sum of % words spelled correctly and % correct writing sequences.

Table K6

Summary Statistics for Scoring Measures for Lord of the Flies

Measure	*M*	*SD*	*Min*	*Max*
Overall (*N* = 80)				
Total words written	388.6	131.8	162	740
Words spelled correctly	378.8	127.8	156	735
Correct writing sequences	343.7	124.2	117	692
Correct minus incorrect writing sequences	274.9	121.8	52	610
% words spelled correctly	97.8	5.8	49.9	100.0
% correct writing sequences	83.2	8.6	59.3	95.2
Production dependent index	1111.1	377.8	440	2167
Production independent index	181.0	11.6	132	195
Year 1 (*n* = 44)				
Total words written	332.9	108.5	162	685
Words spelled correctly	328.5	107.6	156	677
Correct writing sequences	295.6	102.2	117	618
Correct minus incorrect writing sequences	238.1	103.0	52	510
% words spelled correctly	98.6	1.6	92.6	100.0
% correct writing sequences	83.5	9.6	61.9	95.2
Production dependent index	957.0	315.1	440	1980
Production independent index	182.2	10.9	154	195
Year 2 (*n* = 36)				
Total words written	456.8	126.7	239	740
Words spelled correctly	440.1	124.9	239	735
Correct writing sequences	402.6	124.4	222	692
Correct minus incorrect writing sequences	319.9	129.1	106	610
% words spelled correctly	96.8	8.5	49.9	100.0
% correct writing sequences	82.7	7.4	59.3	93.7
Production dependent index	1299.5	365.9	707	2167
Production independent index	179.6	12.4	132	194

Note. Production dependent index is sum of total words written, words spelled correctly, and correct writing sequences. Production independent index is sum of % words spelled correctly and % correct writing sequences.

Appendix L

Intercorrelations Between Scoring Measures by Writing Sample

Table L1

Intercorrelations Between Scoring Measures for "The Stone Boy"[a]

Measure	1	2	3	4	5	6	7	8
1. Total words written	--	1.00*	.96*	.81*	.15	.16	.99*	.17
2. Words spelled correctly		--	.96*	.82*	.19	.18	1.00*	.19
3. Correct writing sequences			--	.94*	.29*	.42*	.98*	.42*
4. Correct minus incorrect writing sequences				--	.43*	.67*	.86*	.66*
5. % words spelled correctly					--	.65*	.21	.73*
6. % correct writing sequences						--	.25*	.99*
7. Production dependent index[b]							--	.26*
8. Production independent index[c]								--

[a] $n = 86$. [b] Sum of total words written, words spelled correctly, and correct writing sequences. [c] Sum of % words spelled correctly and % correct writing sequences.

*$p < .01$.

Table L2

Intercorrelations Between Scoring Methods for Oedipus 1[a]

Method	1	2	3	4	5	6	7	8
1. Total words written	--	1.00*	.97*	.86*	.07	.12	1.00*	.12
2. Words spelled correctly		--	.97*	.88*	.12	.15	1.00*	.15
3. Correct writing sequences			--	.96*	.22	.35*	.99*	.35*
4. Correct minus incorrect writing sequences				--	.37*	.58*	.91*	.58*
5. % words spelled correctly					--	.57*	.14	.69*
6. % correct writing sequences						--	.20	.99*
7. Production dependent index[b]							--	.21
8. Production independent index[c]								--

[a]$n = 77$. [b]Sum of total words written, words spelled correctly, and correct writing sequences. [c]Sum of % words spelled correctly and % correct writing sequences.

*$p < .01$.

Table L3

Intercorrelations Between Scoring Methods for Oedipus 2[a]

Method	1	2	3	4	5	6	7	8
1. Total words written	--	1.00**	.97**	.89**	.21	.08	1.00**	.12
2. Words spelled correctly		--	.98**	.89**	.24*	.10	1.00**	.14
3. Correct writing sequences			--	.95**	.28*	.23*	.99**	.27*
4. Correct minus incorrect writing sequences				--	.32**	.50**	.92**	.52**
5. % words spelled correctly					--	.35**	.24*	.53**
6. % correct writing sequences						--	.14	.98**
7. Production dependent index[b]							--	.18
8. Production independent index[c]								--

[a]$n = 77$. [b]Sum of total words written, words spelled correctly, and correct writing sequences. [c]Sum of % words spelled correctly and % correct writing sequences.

*$p < .05$; **$p < .01$.

Table L4

Intercorrelations Between Scoring Methods for Oedipus 3[a]

Method	1	2	3	4	5	6	7	8
1. Total words written	--	1.00*	.97*	.88*	.09	.11	1.00*	.12
2. Words spelled correctly		--	.98*	.89*	.14	.14	1.00*	.15
3. Correct writing sequences			--	.96*	.21	.31*	.99*	.31*
4. Correct minus incorrect writing sequences				--	.33*	.53*	.91*	.53*
5. % words spelled correctly					--	.57*	.15	.70*
6. % correct writing sequences						--	.19	.99*
7. Production dependent index[b]							--	.19
8. Production independent index[c]								--

[a]$n = 77$. [b]Sum of total words written, words spelled correctly, and correct writing sequences. [c]Sum of % words spelled correctly and % correct writing sequences.

*$p < .01$.

Table L5

Intercorrelations Between Scoring Methods for Lord of the Flies[a]

Method	1	2	3	4	5	6	7	8
1. Total words written	--	.96*	.96*	.81*	-.18	.06	.99*	-.05
2. Words spelled correctly		--	.94*	.82*	.08	.11	.98*	.12
3. Correct writing sequences			--	.94*	-.10	.33*	.98*	.19
4. Correct minus incorrect writing sequences				--	.01	.60*	.87*	.45*
5. % words spelled correctly					--	.25*	-.07	.69*
6. % correct writing sequences						--	.17	.87*
7. Production dependent index[b]							--	.09
8. Production independent index[c]								--

[a]$n = 80$. [b]Sum of total words written, words spelled correctly, and correct writing sequences. [c]Sum of % words spelled correctly and % correct writing sequences.

*$p < .01$.

Appendix M

Summary Statistics by Year for First 100 Words of ECRs

Table M1

Summary Statistics for Scoring Measures for the First 100 Words of "The Stone Boy"

Measure	*M*	*SD*	*Min*	*Max*
Overall (*N* = 86)				
Words spelled correctly	97.8	2.2	87	100
Correct writing sequences	85.4	10.1	57	102
Correct minus incorrect writing sequences	66.6	20.2	13	98
% words spelled correctly	97.8	2.2	87.0	100.0
% correct writing sequences	82.0	9.8	56.4	97.1
Production dependent index	283.2	11.6	248	302
Production independent index	179.8	11.3	146	197
Year 1 (*n* = 45)				
Words spelled correctly	97.9	2.5	87	100
Correct writing sequences	85.5	11.2	57	102
Correct minus incorrect writing sequences	67.2	22.5	13	98
% words spelled correctly	97.9	2.5	87.0	100.0
% correct writing sequences	82.4	10.9	56.4	97.1
Production dependent index	283.4	13.1	248	302
Production independent index	180.3	12.8	146	197
Year 2 (*n* = 41)				
Words spelled correctly	97.7	1.8	92	100
Correct writing sequences	85.3	8.8	64	100
Correct minus incorrect writing sequences	66.0	17.6	23	94
% words spelled correctly	97.7	1.8	92.0	100.0
% correct writing sequences	81.6	8.5	61.0	94.3
Production dependent index	283.0	9.8	260	300
Production independent index	179.3	9.6	158	194

Note. Production dependent index is sum of total words written, words spelled correctly, and correct writing sequences. Production independent index is sum of % words spelled correctly and % correct writing sequences.

Table M2

Summary Statistics for Scoring Measures for the First 100 Words of Lord of the Flies

Measure	*M*	*SD*	*Min*	*Max*
Overall (*N* = 80)				
Words spelled correctly	98.7	2.1	87	100
Correct writing sequences	88.2	9.7	63	105
Correct minus incorrect writing sequences	70.8	19.3	21	103
% words spelled correctly	98.7	2.1	87.0	100.0
% correct writing sequences	83.5	9.1	60.0	98.1
Production dependent index	286.9	11.1	250	305
Production independent index	182.2	10.5	147	198
Year 1 (*n* = 44)				
Words spelled correctly	99.1	1.5	94	100
Correct writing sequences	88.7	10.8	67	105
Correct minus incorrect writing sequences	72.2	21.7	26	103
% words spelled correctly	99.1	1.5	94.0	100.0
% correct writing sequences	84.4	10.4	62.0	98.1
Production dependent index	287.8	11.7	264	305
Production independent index	183.4	11.3	160	198
Year 2 (*n* = 36)				
Words spelled correctly	98.3	2.6	87	100
Correct writing sequences	87.6	8.3	63	101
Correct minus incorrect writing sequences	69.0	16.0	21	93
% words spelled correctly	98.3	2.6	87.0	100.0
% correct writing sequences	82.4	7.4	60.0	94.3
Production dependent index	285.8	10.3	250	301
Production independent index	180.7	9.4	147	194

Note. Production dependent index is sum of total words written, words spelled correctly, and correct writing sequences. Production independent index is sum of % words spelled correctly and % correct writing sequences.

Appendix N

Summary Statistics by Gender for Scoring Measures

Table N1

Summary Statistics by Gender for Scoring Measures for "The Stone Boy"

Measure	*M*	*SD*	*Min*	*Max*
	Female (*n* = 43)			
Total words written	400.4	141.5	169	710
Words spelled correctly	393.5	140.3	169	707
Correct writing sequences	349.8	132.7	117	684
Correct minus incorrect writing sequences	277.6	131.9	35	627
% words spelled correctly	98.2	1.7	93.2	100.0
% correct writing sequences	82.5	9.4	57.7	95.5
Production dependent index[a]	1143.7	410.6	482	2101
Production independent index[b]	180.7	10.5	153	195
	Male (*n* = 43)			
Total words written	336.1	130.7	151	737
Words spelled correctly	329.9	129.3	150	725
Correct writing sequences	287.9	121.8	118	656
Correct minus incorrect writing sequences	221.5	121.2	35	539
% words spelled correctly	98.1	1.7	92.5	100.0
% correct writing sequences	80.7	10.3	58.7	96.5
Production dependent index[a]	953.9	378.3	447	2118
Production independent index[b]	178.8	11.5	153	196

[a]Sum of total words written, words spelled correctly, and correct writing sequences. [b]Sum of % words spelled correctly and % correct writing sequences.

Table N2

Summary Statistics by Gender for Scoring Measures for Oedipus 1

Measure	*M*	*SD*	*Min*	*Max*
	Female (*n* = 40)			
Total words written	124.2	51.9	35	281
Words spelled correctly	122.6	51.5	34	281
Correct writing sequences	114.5	50.0	34	268
Correct minus incorrect writing sequences	96.8	47.6	11	240
% words spelled correctly	98.6	1.4	95.6	100.0
% correct writing sequences	86.3	8.6	55.1	98.0
Production dependent index[a]	361.3	152.6	103	830
Production independent index[b]	184.9	9.3	151	198
	Male (*n* = 37)			
Total words written	110.8	38.8	49	256
Words spelled correctly	108.9	38.6	49	255
Correct writing sequences	99.2	38.2	36	238
Correct minus incorrect writing sequences	80.4	39.5	16	205
% words spelled correctly	98.3	2.4	89.8	100.0
% correct writing sequences	83.8	10.3	58.2	98.7
Production dependent index[a]	318.8	114.5	138	749
Production independent index[b]	182.1	12.0	152	199

[a]Sum of total words written, words spelled correctly, and correct writing sequences. [b]Sum of % words spelled correctly and % correct writing sequences.

Table N3

Summary Statistics by Gender for Scoring Measures for Oedipus 2

Measure	*M*	*SD*	*Min*	*Max*
	Female (*n* = 40)			
Total words written	188.1	78.0	37	394
Words spelled correctly	185.1	76.9	35	386
Correct writing sequences	171.9	72.4	35	350
Correct minus incorrect writing sequences	139.5	69.1	16	293
% words spelled correctly	98.3	1.6	93.0	100.0
% correct writing sequences	84.5	8.8	52.7	97.8
Production dependent index[a]	545.1	226.6	107	1130
Production independent index[b]	182.8	9.3	153	198
	Male (*n* = 37)			
Total words written	145.9	68.0	39	304
Words spelled correctly	143.0	67.6	39	300
Correct writing sequences	133.8	68.8	32	363
Correct minus incorrect writing sequences	109.1	61.0	21	302
% words spelled correctly	97.7	2.6	87.0	100.0
% correct writing sequences	84.0	8.7	61.3	98.1
Production dependent index[a]	422.6	202.9	110	967
Production independent index[b]	181.7	10.1	156	198

[a]Sum of total words written, words spelled correctly, and correct writing sequences. [b]Sum of % words spelled correctly and % correct writing sequences.

Table N4

Summary Statistics by Gender for Scoring Measures for Oedipus 3

Measure	*M*	*SD*	*Min*	*Max*
	Female (*n* = 40)			
Total words written	152.7	60.6	34	261
Words spelled correctly	150.2	59.6	33	258
Correct writing sequences	140.6	56.4	22	254
Correct minus incorrect writing sequences	119.0	52.1	10	239
% words spelled correctly	98.3	2.3	87.0	100.0
% correct writing sequences	86.5	8.0	64.7	97.6
Production dependent index[a]	443.5	175.6	89	755
Production independent index[b]	184.8	9.2	162	198
	Male (*n* = 37)			
Total words written	129.9	54.8	40	248
Words spelled correctly	127.6	54.2	39	244
Correct writing sequences	117.8	52.7	39	240
Correct minus incorrect writing sequences	97.9	50.3	0	207
% words spelled correctly	98.1	2.1	90.7	100.0
% correct writing sequences	84.9	10.1	50.0	97.9
Production dependent index[a]	375.3	160.9	118	732
Production independent index[b]	183.1	11.6	141	197

[a]Sum of total words written, words spelled correctly, and correct writing sequences. [b]Sum of % words spelled correctly and % correct writing sequences.

Table N5

Summary Statistics by Gender for Scoring Measures for Aggregated Oedipus

Measure	*M*	*SD*	*Min*	*Max*
	Female (*n* = 40)			
Total words written	465.1	160.3	106	936
Words spelled correctly	457.8	158.5	102	925
Correct writing sequences	427.0	147.8	91	854
Correct minus incorrect writing sequences	355.4	135.0	68	722
% words spelled correctly	98.3	1.5	93.2	100.0
% correct writing sequences	85.7	6.4	68.8	97.2
Production dependent index[a]	1349.9	464.8	299	2715
Production independent index[b]	184.0	7.0	166	197
	Male (*n* = 37)			
Total words written	386.5	140.5	171	729
Words spelled correctly	379.5	139.6	169	715
Correct writing sequences	350.7	136.5	139	698
Correct minus incorrect writing sequences	287.3	127.5	70	585
% words spelled correctly	98.0	1.9	91.5	99.8
% correct writing sequences	84.4	8.2	65.6	94.9
Production dependent index[a]	1116.7	413.7	479	2085
Production independent index[b]	182.4	9.5	157	194

Note. Aggregated *Oedipus* is the sum of *Oedipus 1*, *Oedipus* 2, and *Oedipus* 3.

[a]Sum of total words written, words spelled correctly, and correct writing sequences. [b]Sum of % words spelled correctly and % correct writing sequences.

Table N6

Summary Statistics by Gender for Scoring Measures for Lord of the Flies

Measure	*M*	*SD*	*Min*	*Max*
	Female (*n* = 41)			
Total words written	414.5	144.2	162	740
Words spelled correctly	399.6	138.5	161	735
Correct writing sequences	361.4	132.9	158	692
Correct minus incorrect writing sequences	281.5	130.4	71	586
% words spelled correctly	97.1	8.2	49.9	100.0
% correct writing sequences	82.1	9.1	59.3	94.7
Production dependent index[a]	1175.4	406.0	481	2167
Production independent index[b]	179.3	13.3	132	195
	Male (*n* = 39)			
Total words written	364.0	115.1	167	665
Words spelled correctly	358.9	115.1	156	664
Correct writing sequences	327.0	114.3	117	654
Correct minus incorrect writing sequences	268.7	114.2	52	610
% words spelled correctly	98.5	1.8	92.6	100.0
% correct writing sequences	84.2	8.2	61.9	95.2
Production dependent index[a]	1049.9	342.8	440	1983
Production independent index[b]	182.6	9.6	154	195

[a]Sum of total words written, words spelled correctly, and correct writing sequences. [b]Sum of % words spelled correctly and % correct writing sequences.

References

Blankenship, C. (1985). Using curriculum-based assessment data to make instructional decisions. *Exceptional Children, 52,* 233-238.

Blankenship, C., & Lilly, S. (1981). *Mainstreaming students with learning and behavior problems: Techniques for the classroom teacher.* New York: Holt, Rinehart, & Winston.

Carver, R.P. (1974). Two dimensions of tests: Psychometric and edumetric. *American Psychologist, 29,* 512-518.

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

Deno, S.L. (1987). Curriculum-based measurement. *Teaching Exceptional Children, 20*, 41-42.

Deno, S.L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention*, *28,* 3-12.

Deno, S.L., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation for written expression. *Exceptional Children, 46,* 368-371.

Deno, S.L., Mirkin, P., & Marston, D., (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (Vol. IRLD-RR-22). University of Minnesota, Institute for Research on Learning Disabilities.

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

Espin, C.A., Scierka, B.J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading & Writing Quarterly, 15,* 5-27.

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

Espin, C., Wallace, T., Campbell, H., Lembke, E.S., Long J.D., & Ticha, R. (2006). *Predicting the success of secondary-second students on state standards tests: Validity and reliability of curriculum-based measures in written expression.* Manuscript submitted for publication.

Espin, C.A., Weissenburger, J.W., & Benson, B.J. (2004). Assessing the writing performance of students in special education. *Exceptionality, 21,* 55-66.

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

Foegen, A., & Deno, S.L. (2001). Identifying growth indicators for low achieving students in middle school mathematics. *Journal of Special Education, 35,* 4-16.

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

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 learning. *American Educational Research Journal, 21*, 449-460.

Fuchs, L.S., & Fuchs, D. (1997). Use of curriculum-based measurement in identifying students with disabilities. *Focus on Exceptional Children, 30(3)*, 1-16.

Fuchs, L.S., Fuchs, D., & Hamlett, C.L. (1989). Effects of instrumental use of curriculum-based measurement to enhance instructional programs. *Remedial and Special Education, 10(2),* 43-52.

Fuchs, L.S., Fuchs, D., Hosp, M.K., & Hamlett, C.L. (2003). The potential for diagnostic analysis within curriculum-based measurement. *Assessment for Effective Intervention, 28,* 13-22.

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

Gansle, K.A., Noell, G.H., VanDerHeyden, A.M., Slider, N.J., Hoffpauir, L.D., & Whitmarsh, E.L. (2004). An examination of the criterion validity, and time cost of alternative measures for curriculum-based measurement in writing. *School Psychology in the Schools, 41*, 291-300.

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

Gickling, E.E., Shane, R.L., & Croskery, K.M. (1989). Assuring math success for low-achieving high school students through curriculum-based assessment. *School Psychology Review, 18,* 344-356.

Gickling, E.E., & Thompson, V.P. (1985). A personal view of curriculum-based assessment. *Exceptional Children, 52,* 205-218.

Hintze, J.M., & Silberglitt, B. (2005). Longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high stakes testing. *School Psychology Review, 34,* 372-386.

Howell, K.W., & Morehead, M.K. (1987*). Curriculum-based evaluation for special and remedial education.* Columbus, OH: Charles Merrill.

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

Ketterlin-Geller, L.R., McCoy, J.D., Twyman, T., & Tindal, G. (2006). Using a concept maze to assess student understanding of secondary content. *Assessment for Effective Intervention, 31(2),* 39-50.

Knoff, H.M., & Dean, K.R. (1994). Curriculum-based measurement of at-risk students' reading skills: A preliminary investigation of bias. *Psychological Reports, 75,* 1355-1360.

Kranzler, J.H., Miller, M.D., & Jordan, L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly, 14,* 327-342.

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

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

Marston, D., & Magnusson, D. (1988). Curriculum-Based Measurement: District level implementation. In J. L. Graden, J. E. Zins, & M. J. Curtis (Eds.), *Alternative*

educational delivery systems: Enhancing instructional options for all students (pp. 137-172). Washington, D.C.: National Association of School Psychologists.

Marston, D., Mirkin, P.K., & Deno, S.L. (1984). Curriculum-based measurement: An alternative to traditional screening, referral, and identification. *Journal of Special Education, 18,* 109-118.

Maryland State Department of Education (n.d.). *Helpful tools for the high school assessments.* Retrieved September 12, 2007, from http://hsaexam.org/

Maryland State Department of Education. (2005). What is the high school assessment program? In *How do we test what students have learned? (9-12*). Retrieved September 12, 2007, from http://www.mdk12.org/mspp/high_school/what_is/index/html

Maryland State Department of Education. (2007, June). *Maryland High School Assessment 2006 technical report.* Retrieved September 12, 2007, from http://www.hsaexam.org/moreinfo.html

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

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

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

Parker, R., Tindal, G., & Hasbrouck, J. (1991b). Progress monitoring with objective measures of writing performance for students with mild disabilities. *Exceptional Children, 58,* 61-73.

School Improvement in Maryland. (n.d.). Using the Core Learning Goals: English. In *Teaching and Learning: Reading/ELA*. Retrieved September 18, 2007, from http://www.mdk12.org/instruction/clg/english/goal2.html

Shinn, M. (1995). Best practices in curriculum-based measurement and its use in a problem-solving model. In J.Grimes & A. Thomas (Eds). *Best practices in school psychology III* (pp. 547-568). Silver Spring, MD: National Association of School Psychologists.

Shinn, M.R., Rosenfield, S., & Knutson, N. (1989). Curriculum-based assessment: A comparison of models. *School Psychology Review, 19,* 299-316.

Shinn, M.R., Ysseldyke, J.E., Deno, S.L., & Tindal, G.A. (1986). A comparison of differences between students labeled learning disabled and low achieving on measures of classroom performance. *Journal of Learning Disabilities*, *19,* 545-552.

Swanson, Gwenyth. (2007). *How do scorers assess HSA?* Retrieved September 23, 2007, from http://mdk12.org/mspp/high_school/look_like/hsa_scored.html

Tindal, G., & Nolet, V. (1995). Curriculum-based measurement in middle and high schools: Critical thinking skills in content areas. *Focus on Exceptional Children, 27*, 1-22.

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

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

U.S. Department of Education. Institute of Education Sciences. National Center for Education Statistics. *The Nation's Report Card: Writing 2002*, NCES 2003-529, by H.R. Persky, M.C. Daane, and Y. Jin. Washington, DC: 2003.

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

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

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