VALIDITY OF TWO GENERAL OUTCOME MEASURES OF SCIENCE AND SOCIAL STUDIES ACHIEVEMENT

Paul Mooney  
Special Education Programs  
Louisiana State University  
213 Peabody Hall, USA

Renée E. Lastrapes  
University of Houston  
UH-Clear Lake Box 55  
2700 Bay Area Blvd. Houston, Texas 77058

Amanda M. Marcotte  
University of Massachusetts Amherst  
S154 Furcolo  
Amherst, MA 01003

Amy Matthews, B. S.  
Louisiana State University  
13046 LA-73, Geismar, LA 70734

Abstract

The present research expanded validity findings for a structured formative assessment measure of content learning that was administered online and known as critical content monitoring. The study also evaluated the potential for additional measures, including sentence verification technique and written retell, to explain variance in student achievement in science and social studies classrooms. Participants were fifth-grade students (N=51) enrolled in a public primary school in the southeastern U.S. Three predictor variables (i.e. critical content monitoring, sentence verification technique and written retell) were correlated with content test scores from the nationally representative standardized achievement test (i.e. Stanford Achievement Test-Tenth Edition abbreviated online form) and a statewide accountability test. Pearson correlations for critical content monitoring and the Stanford tests across science (r=.55) and social studies (r=.63) were moderately strong and similar in magnitude with other reported correlations for academic language measures in the literature. Correlations for critical content monitoring were descriptively larger than those between the standardized tests and sentence verification technique and written retell. Commonality
analyses indicated that both critical content monitoring and sentence verification technique added unique variance to explanatory models. Limitations and implications were discussed.

**Keywords:** structured formative assessment, general outcome measurement, content courses.

Structured formative assessment has long been part of the fabric of special education, both in the profession’s practice and promise (Jenkins, & Fuchs, 2012). Since Dr. Stanley Deno’s development and introduction of curriculum-based measurement (Deno, 1985), special education teachers have had the capacity to monitor their students’ progress toward individualized education programme goals. Curriculum-based measurement has been implemented and evaluated primarily in the elementary grades and in the area of reading skill development. However, from its inception, assessment procedures have also been available in the areas of beginning math, spelling, and writing skills. What did not exist initially were procedures to evaluate student performance and progress in content areas such as social studies and science. It is the content areas to which this research applies.

Inquiry addressing the efficacy of structured formative assessment techniques to document student achievement and growth in social studies and science content is still in its infancy. Early on, researchers such as Espin and Foegen (1996) wondered whether available reading measures such as oral reading fluency and maze could also serve to document performance in content areas. They found that a measure of content vocabulary was a descriptively stronger correlate than the reading measures. From that beginning has evolved a more focused evaluation of academic content-driven measures, including tests of the validity of critical content monitoring (Mooney, McCarter, Russo, & Blackwood, 2013), an online-administered and scored structured formative assessment tool, that is the focus of the present study. What follows is a description of critical content monitoring and a rationale for the following two research questions:

1. What were the correlations with nationally-normed standardized achievement and statewide accountability tests in science and social studies content for critical content monitoring, sentence verification technique and written retell?
2. What was the incremental validity of adding measures of reading comprehension, using sentence verification technique and/or written retell, to content-focused achievement models that included critical content monitoring?

**Critical content monitoring**

Critical content monitoring was originally developed using a curriculum sampling approach (Fuchs, 2004). Probes were created by sampling the corpus of content vocabulary across a grade-level curriculum in order to ensure that all important content was included across all forms. As a general outcome measure, its aim has been to serve as an index of content learning at both a point in time and over time in order to operate formatively and improve instructional decision-making. Administration of critical content monitoring probes involves students reading definitions of key grade-level science or social studies vocabulary at a secure learning management system link and choosing the correct answer from a list of terms. Students generally have up to 5 minutes to answer 20 questions.

Academic vocabulary was utilized as an indicator of performance because it serves as communicative currency (Alexander, n.d.) in content courses. That is, the words, phrases, and concepts of the subject matter form the content of activities in the classroom. Success in the content classroom occurs through relevant and meaningful employment of academic vocabulary. Demonstration of academic language’s robustness is evident in the stronger correlations with a relevant criterion for vocabulary matching over competing measures including oral reading fluency, maze and a writing measure (Espin, & Foegen, 1996; Mooney, McCarter, Schraven, & Callicoatte, 2013).

Critical content monitoring is an online adaptation of vocabulary matching (Espin & Deno, 1994-1995). Its research has addressed technical concerns of a single probe score. Mooney, McCarter, Russo et al. (2013) assessed the criterion validity for a collection of 20 science probes in relation to a statewide accountability test for a sample of generally high-performing fifth-grade students. The results indicated moderate correlations for the 20 probes ($r=.36$ - $.55$). Mooney, McCarter, Russo, Blackwood (2014) extended the criterion validity findings, demonstrating a moderately strong correlation (.67) between a critical content monitoring social studies probe and the statewide content test. The social studies correlation was descriptively larger than the science correlation and comparable in magnitude to previous vocabulary matching findings ($rs .64$ to $.70$; Espin, Shin, & Busch, 2005; Mooney, McCarter, & Schraven et al., 2013).
Rationale for study

The predictor measures in the present study were either originally designed as general outcome measures of learning or more recently adapted for that purpose. General outcome measurement (GOM) is one of two instructionally relevant measurement models (Fuchs & Deno, 1991). An alternative to subskill mastery measurement, its original goal was to establish tests that: (a) evidenced reliability and validity, and (b) assisted teachers in planning better instructional programmes and evaluating instructional programme success. These findings enhance the larger literature in content-oriented structured formative assessment, including studies that have addressed the performance and progress tenability of vocabulary matching (e.g., Espin, Lembke, Hampton, Seo, & Zukowski, 2013; Mooney, McCarter, & Schraven et al., 2013).

General outcome measures of academic language, including critical content monitoring, vocabulary matching and key vocabulary (Vannest, Parker, & Dyer, 2011) have the potential to predict achievement and inform instructional decision-making. Critical content monitoring, for example, was intended as a measure of science or social studies course learning by content teachers in the upper elementary and secondary school grades. For teachers, academic language is an alterable variable (Bloom, 1980) that is particularly pertinent to their interaction with struggling learners. Still, more research is warranted as GOM-type assessment tools are created to reflect broader academic domains. Sentence verification technique and written retell were evaluated as potential general outcome measures of reading comprehension by Marcotte and Hintze (2009). General outcome measures, particularly in reading, have been demonstrated to be effective tools to predict student achievement and inform teacher decision-making (Stecker, Fuchs, & Fuchs, 2005; Wallace, Espin, McMaster, Deno, & Foegen, 2007). Moreover, they have been instrumental in the implementation of responsiveness-to-intervention (RTI) frameworks.

The aforementioned research questions relate to what Fuchs (2004) termed Stage 1 viability of the static score, and Deno and Fuchs (1987) categorized as technical adequacy questions in their instrument development matrix. The first question addressed the need to extend validity research for critical content monitoring. The measure has validity findings comparing scores with a statewide accountability test across both science and social studies content (Mooney, McCarter, & Russo et al., 2013, 2014). The first research question extended criterion validity evidence by comparing critical content monitoring scores with those of a standardized measure of content knowledge. Generalization concerns exist because critical monitoring comparisons to date have been made with state-specific accountability tests. The present inquiry was the first
comparison between scores of critical content monitoring and a standardized, nationally-recognized test of science and social studies achievement. It was hypothesized that correlations with a nationally-representative test would be comparable to those with statewide accountability tests.

The second research question addressed the complexity of knowledge and skills that may be relevant for measurement in a formative assessment model. Indicators of academic language, such as vocabulary matching, critical content monitoring and key vocabulary, are not the only variables that predict criterion achievement so other indicators of success in content areas should be evaluated. That is, while students are expected to master academic language, they are also expected to read and comprehend instructional texts and relevant materials presented in class (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) as well as summarize and apply what they learn and read in written form.

Marcotte and Hintze (2009) evaluated a series of reading comprehension explanatory models that paired oral reading fluency with a number of potential GOM comprehension tools, including maze, retell fluency, sentence verification technique and written retell. Correlations for all of the GOM predictors with a criterion were moderately strong, ranging from .46 to .67. Multiple-regression analyses indicated that a collection of four of the measures (i.e. oral reading fluency, maze, sentence verification technique and written retell) accounted for about 57% of the variability in overall reading ability. Moreover, the maze, sentence verification technique and written retell measures were all found to contribute significantly to the overall explanatory model after controlling for oral reading fluency. One takeaway from the study was that four measures together proved to be a better predictive mechanism than one measure alone.

In the present study three measures targeting academic language and comprehension were evaluated to determine whether there would be a greater share of the variability in content achievement for multiple measures of content achievement beyond that provided for by critical content monitoring. Potential GOM tools, sentence verification technique and written retell, were chosen over oral reading fluency and maze because of their potential value in describing variance in content achievement. Moreover, previous research demonstrated that a measure of academic language was the more strongly correlated instrument with a criterion when comparisons were made either descriptively or directly (Espin, Foegen, 1996; Mooney, McCarter, & Schraven et al., 2013), making it a preferable capstone task. It was hypothesized that a multi-element assessment system would be a better predictor than a single-measure system.
Method
Participants and setting

Participants in the institutional review board-approved study were fifth-graders (N = 51) in a single public prekindergarten to fifth grade school in south Louisiana whose parents consented and who they assented to be involved in the study. As a whole, participants were 11.1 years old (SD = .5) at the time of testing, 68.6% female (n = 35), 66.7% Caucasian (n = 34), and 68.6% full-pay lunch status (n = 35). 96% of the participants (n = 49 in both cases) scored at the ‘basic’ level of proficiency on the statewide accountability content tests (which would be considered passing if science and social studies test scores were categorized in the same manner as the English language arts and math tests were in Louisiana). Statewide, 43% and 47% of fifth-graders scored at ‘basic’ in science and social studies, respectively, on the state test. Assenting participants were 40.8% of an entire fifth-grade class.

Measures

Five measures were compared in the present study. The criterion measures were the science and social studies content tests of the online abbreviated Stanford Achievement Tests-Tenth Edition (SAT-10; Pearson Education, n.d.) and the integrated Louisiana Educational Assessment Programme (iLEAP; LDE, n.d., a). The predictor variables were critical content monitoring, sentence verification technique and written retell.

SAT-10. The abbreviated form of the online SAT-10 is a standardized, norm-referenced achievement test battery that measures reading, mathematics, spelling, language, listening, science and social studies performance for students in kindergarten through the 12th grade. For this study, only grade-level science and social studies tests were administered. Publishers described the content tests as reflecting current practice and research and aligned with national and state content standards. The science test assessed knowledge of life, physical and earth sciences as well as science as inquiry. The social studies test assessed knowledge of history, geography, political science and economics. Each abbreviated battery content test consisted of 30 multiple-choice questions and was untimed. The test-derived scaled score was used in the present study. The scaled score is vertically equated across each subject test, reportedly allowing for the tracking of performance across grades (Pearson Education, n.d.). For the Spring 2013 testing period, the sample’s average science score was 658 (range 602-726), placing the collective group at the 63rd percentile nationally. The
average social studies score of 655 (range 597-713) was reported at the 59th percentile. Two online Buros Institute *Mental Measurements Yearbook* reviewers (Carney, n.d.; Morse, n.d.) provided support for the use of SAT-10 as a whole in measuring achievement in public schools. Reviewers described alternate-form reliability and content validity evidence for SAT-10. In the present study, criterion validity was evidenced by moderately strong correlations between the SAT-10 and *iLEAP* content tests, with .64 and .69 linear relations, both \( p < .01 \), for science and social studies, respectively.

*iLEAP grade 5 criterion-referenced test.* The stated purpose of *iLEAP* is measurement toward Louisiana’s academic standards in English language arts, math, science and social studies (LDE, n.d., a) for all students in grades 3, 5, 6, 7 and 9. The science and social studies tests included multiple-choice questions, were untimed and administered on different days. Fifth grade science content strands included science as inquiry, physical, life, earth, space and environmental science, with test questions addressing all strands. Social studies content strands included geography, civics, economics and history, with test questions addressing only the geography and history strands (LDE, n.d., a). Achievement level descriptors were unsatisfactory, approaching basic, basic, mastery and advanced. Technical adequacy data for the *iLEAP* fifth grade tests were accessed from the LDE website. Cronbach's alpha levels of .85 for science and .82 for social studies were reported as reliability evidence of the 2010 test’s internal consistency (LDE, n.d., b). State-provided validity data were described in terms of a content validity process that was not delineated (LDE, n.d., b).

**Critical content monitoring.** The content-focused general outcome measure described earlier evolved from procedures previously outlined in Espin, Busch, Shin, Kruschwitz (2001). Terms in each researcher-created probe were randomly selected from the full body of content terms collected from textbook glossary sections and reviewed for legitimacy by a small group of practicing teachers recommended by the first author as both content knowledgeable and effective teachers. The list of terms was organized by curricular unit. Each probe included terms from each unit. To generate each probe, the number of terms per unit was determined by calculating the proportion of the year’s curriculum that was devoted to each unit in the state pacing guide and then multiplying that proportion by the number of questions in each probe. Twenty terms and accompanying definitions were entered into an online learning management system (Moodle, n.d.) in a multiple-choice format. Alternate form reliability correlations for 20 parallel probes had a mean correlation of .55 (range .21 to .73; SD = .09) (Mooney, McCarter, & Russo et al., 2013). Criterion validity findings were previously reported.
Sentence verification technique. Sentence verification technique was reportedly created as a reading comprehension assessment method (Royer et al., 1979). The measure consisted of reading passages that were followed by sentences that test takers read and responded to after reading the passage and without access to the passage content. Sentences were developed from the passages that test takers read and were one of the following type: (a) originals, or exact copies of passage sentences; (b) paraphrases, or similar-meaning sentences with built-in word changes; (c) meaning changes, or similar sentences with slight changes in words to alter meaning; and (d) distractors, or similar topic sentences that differ in both meaning and wording from the passage. For the present study, passage content was adapted from approved grade-level science and social studies texts. Passages were split between science and social studies content and introduced alternately, with 16 sentences following each passage. An examinee read each passage and then responded to the sentences with a yes (i.e. the meaning of the sentence is the same as that of the passage) or no (i.e. the meaning is different) response. Criterion validity correlations with standardized test measures, including the SAT, have ranged from .50 to .73; reliability for four-passage probes have ranged from .70 to .80 (Royer, 2004).

Written retell. The format for written retell was identical to that utilized in Marcotte and Hintze (2009). That is, students were asked to read a 750-word passage silently for 5 minutes. Then, the passage was removed and students were asked to write down all that they could remember about the passage. Students had 5 minutes to respond to the initial prompt, which included reminders of the task requirement offered periodically during that time span. Scores for written retell consisted of the number of unique content words written by each student. As in Marcotte and Hintze, content words were defined as “distinct proper and common nouns, verbs, adjectives, and adverbs contained in the passage or synonymous with those in the passage” (p. 322). The list of content words was developed by the first author based upon a reading of the passage. A criterion validity correlation for written retell with a standardized achievement test was reported at .57 (Marcotte & Hintze). The written retell probe also demonstrated .56 and .59 correlations with oral reading fluency and maze, respectively.

Procedures

Testing took place during science class in mid-May 2013, near the end of the school year and about six weeks after statewide accountability testing. Originally, test administration was designed in a counter-balanced arrangement by class section in order to address order effects. However, technical difficulties
encountered on the first occasion of SAT-10 online testing resulted in an alteration of the original schedule, with SAT-10 testing taking place last for all participants. For SAT-10 and critical content monitoring, students logged on to secure sites under the supervision of the first author. For sentence verification technique and written retell, testing was directed by the first and fourth authors, with the latter the classroom science teacher. For the statewide accountability tests, the classroom teacher oversaw administration.

**Interscorer agreement**

Sentence verification technique and written retell were independently scored by the first and second authors. The two scores were entered into a database and checked to ensure accurate data entry. Initial data analyses included the calculation of agreement proportions for all participant total scores. An agreement occurred when both scorers reported the same total score per individual probe. Agreement proportions were 100% for sentence verification technique and 85.7% for written retell. The first scores from the first author were used for the analyses. No interscorer reliability actions were taken for critical content monitoring beyond checking the online system to ensure that the right answer choice accompanied each stem. SAT-10 scores were provided by the publishing company whereas iLEAP scores were provided by the teacher.

**Analysis**

Analyses addressed criterion and incremental validity questions. Criterion validity for each of the three predictor variables was assessed using the correlation and the 95% confidence interval (CI) between each probe and a relevant criterion measure. All variables were assessed and deemed normally distributed using the Shapiro-Wilk statistic. Incremental validity was assessed using sequential multiple linear regression and commonality analysis. Separate sequential or hierarchical multiple regression analyses were performed for science and social studies content areas. Sequential regression was used to determine if information from additional predictor variables improved prediction of the criterion measure after the effects of previous variables entered had been statistically eliminated (Tabachnick & Fidell, 2013). In order to determine how much more variance was accounted for by each of the comprehension measures critical content monitoring was entered first, followed by sentence verification technique and then written retell. Finally, commonality analyses were conducted to determine the amount of variation in the criterion
variables (i.e. SAT-10 science and social studies) accounted for by each of the predictor variables. Commonality analysis is designed to quantify the unique explanatory power of each predictor as well as the explanatory power that is common to all possible combinations of the predictors (Zientek & Thompson, 2006). The commonality formulas for Excel were obtained from Warne (2011).

Results
Criterion validity

Table 1 displays sample means, distributions and 95% CIs of state and national content tests and each of the three predictors. While skewness and kurtosis were evident for the predictors, normality checks showed that all but the critical content monitoring science scores were normally distributed. Table 2 displays correlations of all assessments with 95% CIs. Correlations and 95% CIs between the predictor and criterion variables were in the low (i.e. <.3) to moderate (i.e. .3 to .7) range. Critical content monitoring was most highly correlated with both content tests, followed next by sentence verification technique and then written retell. Social studies correlations were descriptively greater in magnitude than those in science. Each of the correlations for critical content monitoring and sentence verification technique and the criterion measures were significantly different from zero ($p < .01$). Sentence verification technique showed stronger correlations with the social studies tests than science, and WRT was only significantly correlated with the state social studies test.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Range</th>
<th>Mean</th>
<th>SD</th>
<th>95% CI</th>
<th>Skew-ness</th>
<th>Kurto-sis</th>
</tr>
</thead>
<tbody>
<tr>
<td>iLEAP Science</td>
<td>51</td>
<td>276-454</td>
<td>360.6</td>
<td>36.8</td>
<td>350, 371</td>
<td>.33</td>
<td>.62</td>
</tr>
<tr>
<td>iLEAP Social Studies</td>
<td>51</td>
<td>281-424</td>
<td>347.3</td>
<td>31.7</td>
<td>338, 356</td>
<td>.29</td>
<td>-.01</td>
</tr>
<tr>
<td>SAT-10 Science</td>
<td>51</td>
<td>602-726</td>
<td>658.5</td>
<td>25.6</td>
<td>651, 666</td>
<td>.15</td>
<td>.29</td>
</tr>
<tr>
<td>SAT-10 Social Studies</td>
<td>46</td>
<td>586-713</td>
<td>654.0</td>
<td>27.7</td>
<td>646, 662</td>
<td>.06</td>
<td>.17</td>
</tr>
<tr>
<td>CCM Science</td>
<td>49</td>
<td>8-20</td>
<td>15.73</td>
<td>3.05</td>
<td>14.86, 16.61</td>
<td>-.80</td>
<td>.02</td>
</tr>
<tr>
<td>CCM Social Studies</td>
<td>50</td>
<td>3-18</td>
<td>11.34</td>
<td>4.34</td>
<td>10.08, 12.60</td>
<td>-.08</td>
<td>-1.08</td>
</tr>
<tr>
<td>SVT</td>
<td>50</td>
<td>29-59</td>
<td>47.36</td>
<td>6.54</td>
<td>45.50, 49.22</td>
<td>-.32</td>
<td>.34</td>
</tr>
<tr>
<td>WRT</td>
<td>49</td>
<td>11-42</td>
<td>26.92</td>
<td>7.86</td>
<td>24.66, 29.18</td>
<td>.39</td>
<td>-.78</td>
</tr>
</tbody>
</table>

Note. CCM = critical content monitoring; SVT = sentence verification technique; WRT = written retell; iLEAP = integrated Louisiana Educational Assessment Program; SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form.
Table 2

Correlations and 95% Confidence Intervals Among all Predictor and Criterion Measures

<table>
<thead>
<tr>
<th>Test</th>
<th>(i)LEAP Science</th>
<th>SAT Science</th>
<th>SAT Social Studies</th>
<th>CCM Science</th>
<th>CCM Social Studies</th>
<th>SVT</th>
<th>WRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)LEAP Science</td>
<td>0.66** [0.47, 0.79]</td>
<td>0.64** [0.44, 0.78]</td>
<td>0.57** [0.35, 0.73]</td>
<td>0.51** [0.27, 0.69]</td>
<td>0.47** [0.22, 0.66]</td>
<td>0.46** [0.21, 0.65]</td>
<td>0.24 [0.03, 0.48]</td>
</tr>
<tr>
<td>(i)LEAP Social Studies</td>
<td>0.49** [0.25, 0.68]</td>
<td>0.69** [0.54, 0.81]</td>
<td>0.61** [0.40, 0.76]</td>
<td>0.66** [0.47, 0.79]</td>
<td>0.49** [0.25, 0.68]</td>
<td>0.29* [0.02, 0.52]</td>
<td></td>
</tr>
<tr>
<td>SAT Science</td>
<td></td>
<td>0.51** [0.27, 0.69]</td>
<td>0.55** [0.32, 0.72]</td>
<td>0.43** [0.18, 0.63]</td>
<td>0.49** [0.25, 0.68]</td>
<td>0.16 [0.08, 0.42]</td>
<td></td>
</tr>
<tr>
<td>SAT Social Studies</td>
<td></td>
<td></td>
<td>0.65** [0.26, 0.70]</td>
<td>0.63** [0.42, 0.77]</td>
<td>0.59** [0.36, 0.75]</td>
<td>0.28 [0.09, 0.53]</td>
<td></td>
</tr>
<tr>
<td>CCM Science</td>
<td></td>
<td></td>
<td></td>
<td>0.60** [0.38, 0.75]</td>
<td>0.40** [0.13, 0.61]</td>
<td>0.26 [0.05, 0.49]</td>
<td></td>
</tr>
<tr>
<td>CCM Social Studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.41** [0.15, 0.62]</td>
<td>0.25 [0.03, 0.49]</td>
<td></td>
</tr>
<tr>
<td>SVT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.22 [0.06, 0.46]</td>
<td></td>
</tr>
</tbody>
</table>

Note. ** Correlation is significant at the .01; * correlation is significant at .05 level. CCM = critical content monitoring; SVT = sentence verification technique; WRT = written retell; \(i\)LEAP = integrated Louisiana Educational Assessment Program; SAT = Stanford Achievement Test-Tenth Edition, abbreviated form.

**Incremental validity**

Based on the results of the correlational analysis, the predictors were entered into the sequential regression model with critical content monitoring first, sentence verification technique second and written retell third. Data were analyzed for influential points and multicollinearity. Because two observations showed values of Cook’s D that were greater than 1, regression analyses were conducted with and without them and results did not change appreciably; therefore, the observations were retained in the data set. Formal tests for multicollinearity revealed that none was detected. Residual plots indicated homoscedasticity as well as linear relationships among all predictor and criterion variables.
Tables 3 and 4 display results of the regression analyses for the three predictor variables and the criterion SAT-10. Findings indicated that the greatest amount of variation was associated with critical content monitoring and sentence verification technique. Written retell was not a significant predictor either science or social studies, with adjusted $R^2$ decreasing for both SAT-10 test scores with its inclusion.

**Table 3**

**Sequential Regression for SAT-10 Science with Predictors Entered in the Order Presented**

<table>
<thead>
<tr>
<th>Predictors</th>
<th>$R$</th>
<th>$R^2$</th>
<th>Adj. $R^2$</th>
<th>$R^2$ change</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCM</td>
<td>0.54</td>
<td>0.30</td>
<td>0.28</td>
<td>0.30</td>
<td>4.83 .000</td>
<td>3.75 .002</td>
<td>3.79 .002</td>
</tr>
<tr>
<td>CCM, SVT</td>
<td>0.61</td>
<td>0.37</td>
<td>0.35</td>
<td>0.08</td>
<td>1.21 .023</td>
<td>1.22 .024</td>
<td></td>
</tr>
<tr>
<td>CCM, SVT, WRT</td>
<td>0.61</td>
<td>0.37</td>
<td>0.33</td>
<td>0.00</td>
<td>-0.08 .855</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form; CCM = critical content monitoring; SVT = sentence verification technique; WRT = written retell.

**Table 4**

**Sequential Regression for SAT-10 Social Studies with Predictors Entered in the Order Presented**

<table>
<thead>
<tr>
<th>Predictors</th>
<th>$R$</th>
<th>$R^2$</th>
<th>Adj. $R^2$</th>
<th>$R^2$ change</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCM</td>
<td>0.62</td>
<td>0.38</td>
<td>0.36</td>
<td>0.38</td>
<td>3.64 .000</td>
<td>2.73 .000</td>
<td>2.65 .000</td>
</tr>
<tr>
<td>CCM, SVT</td>
<td>0.72</td>
<td>0.52</td>
<td>0.50</td>
<td>0.14</td>
<td>1.76 .001</td>
<td>1.72 .001</td>
<td></td>
</tr>
<tr>
<td>CCM, SVT, WRT</td>
<td>0.73</td>
<td>0.53</td>
<td>0.49</td>
<td>0.01</td>
<td>0.23 .452</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form; CCM = critical content monitoring; SVT = sentence verification technique; WRT = written retell.

Table 5 displays results of the commonality analyses. For science the regression effect was most influenced by critical content monitoring (14.8%) and sentence verification technique (7.3%). There was no measurable variance that was uniquely attributable to written retell. When examined in combination, critical content monitoring and sentence verification technique accounted for a third of the variation in the criterion variable (14.8 + 7.3 + 12.8 = 34.9%).
When examined in combination with critical content monitoring, written retell appeared to have a suppressor effect on critical content monitoring, as evidenced by the negative coefficient. Table 5 also indicates that for SAT-10 social studies scores, the regression effect was most influenced uniquely by critical content monitoring (17.3%) and sentence verification technique (11.3%). The variance that was unique to written retell again appeared to act as a suppressor variable (-1.6% of the total). When examined in combination, critical content monitoring and sentence verification technique accounted for 44.9% of the variation in SAT-10 social studies (17.3 + 11.3 + 16.3).

### Table 5

<table>
<thead>
<tr>
<th></th>
<th>SAT Science</th>
<th>CCM</th>
<th>SVT</th>
<th>WRT</th>
<th>R² Partition</th>
<th>SAT Social Studies</th>
<th>CCM</th>
<th>SVT</th>
<th>WRT</th>
<th>R² Partition</th>
</tr>
</thead>
<tbody>
<tr>
<td>U₁</td>
<td>14.8%</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>14.8%</td>
<td>U₁</td>
<td>17.3%</td>
<td>--</td>
<td>--</td>
<td>17.3%</td>
</tr>
<tr>
<td>U₂</td>
<td>--</td>
<td>7.3%</td>
<td>--</td>
<td>--</td>
<td>7.3%</td>
<td>U₂</td>
<td>--</td>
<td>11.3%</td>
<td>--</td>
<td>11.3%</td>
</tr>
<tr>
<td>U₃</td>
<td>--</td>
<td>--</td>
<td>0.0%</td>
<td>0.0%</td>
<td>--</td>
<td>U₃</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>-1.6%</td>
</tr>
<tr>
<td>C₁₂</td>
<td>12.8%</td>
<td>12.8%</td>
<td>--</td>
<td>--</td>
<td>12.8%</td>
<td>C₁₂</td>
<td>16.3%</td>
<td>16.3%</td>
<td>--</td>
<td>16.3%</td>
</tr>
<tr>
<td>C₁₃</td>
<td>-1.0%</td>
<td>--</td>
<td>-1.0%</td>
<td>--</td>
<td>-1.0%</td>
<td>C₁₃</td>
<td>1.7%</td>
<td>--</td>
<td>1.7%</td>
<td>1.7%</td>
</tr>
<tr>
<td>C₂₃</td>
<td>--</td>
<td>0.1%</td>
<td>0.1%</td>
<td>0.1%</td>
<td>--</td>
<td>C₂₃</td>
<td>3.1%</td>
<td>3.1%</td>
<td>3.1%</td>
<td>3.1%</td>
</tr>
<tr>
<td>C₁₂₃</td>
<td>3.4%</td>
<td>3.4%</td>
<td>3.4%</td>
<td>3.4%</td>
<td>--</td>
<td>C₁₂₃</td>
<td>4.4%</td>
<td>4.4%</td>
<td>4.4%</td>
<td>4.4%</td>
</tr>
<tr>
<td>Sum = r²</td>
<td>30.0%</td>
<td>23.6%</td>
<td>2.5%</td>
<td>--</td>
<td>39.7%</td>
<td></td>
<td>35.1%</td>
<td>7.6%</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Sum = R²</td>
<td>37.4%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>52.5%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form; CCM = critical content monitoring; SVT = sentence verification technique; WRT = written retell. U = Unique; C = Common; 1 = CCM; 2 = SVT; 3 = WRT.

### Discussion

Providing evidence of test score validity is a vital consideration in test development (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999). It is also an important notion for special educators who are charged with individualizing educational programming and ensuring that those unique practices are based on research to the extent practicable. The present study
offered further evidence of criterion validity for critical content monitoring as well as initial evidence of validity for a package of potential GOM instruments for formative use in public school science and social studies classrooms.

Critical content monitoring scores correlated moderately in magnitude (i.e. .3 to .7) in science and social studies to scores from the nationally-recognized SAT-10, which was expected and necessary for consideration of the generalizability of these findings. The science correlation of .55 (95% CI; .32, .72) was within the .36 to .55 range for linear relations with a statewide test previously reported in Mooney, McCarter, Russo et al. (2013). The moderate magnitude of correlation between critical content monitoring and state level accountability test (i.e. .51) was replicated as well (see Table 2). The social studies correlation with SAT-10 of .63 (95% CI; .42, .77) was descriptively higher than the science correlation and comparable to the .68 linear relation with a state accountability test reported in Mooney et al. (2014) and .66 in this study (see Table 2). A pattern of critical content monitoring relations that were descriptively higher in social studies than in science was maintained across national and state assessments. Further evidence of the validity of the critical content monitoring score was indicated by its explanatory power. In regression models identifying the amount of variance accounted for by predictor variables, critical content monitoring was the strongest predictor for both science and social studies tests of the SAT-10. Moreover, commonality analyses indicated that 15% to 17% of the variance in models was uniquely explained by critical content monitoring.

Findings from the present study also identified sentence verification technique as a potential indicator of student performance in the content area courses, with 7% to 11% of the variance explained by sentence verification technique alone and roughly 35% to 45% explained by critical content monitoring and sentence verification technique combined. Written retell was not a significant contributor to the regression effect and showed little common variance with the criterion variable. Both previously had been evaluated as potential measures of reading comprehension by Marcotte and Hintze (2009). For sentence verification technique, which was previously found to have moderate correlations with criteria and explanatory power in models of reading comprehension (Marcotte, Hintze), scores were also moderately correlated with science and social studies content test scores of the SAT-10, with linear relations of .49 (95% CI; .25, .67) in science and .59 (95% CI; .37, .75) in social studies. These results provide evidence that critical content vocabulary and reading comprehension are meaningful predictors of achievement in science and social studies as measured by both state and nationally represented achievement tests.
with sentence verification technique a strong potential candidate as an effective
general outcome measure of content knowledge.

For written retell, neither of its correlations with the SAT-10 was moderate
in magnitude or statistically significant in this study. Findings contrast those
of Marcotte and Hintze (2009), who found that written retell was a significant
predictor of models of reading comprehension. The disparate findings for
written retell may have resulted from the fact that written retell is scored by
the number of meaningful words recalled and written from the passage, which
in the context of science, students may have less familiarity with the vocabulary
and thus have been less likely to recall them and write them down.

**Limitations**

Three primary limitations were noted in the present study. First, due to
technical difficulties during the original implementation of the research plan,
the order of testing was altered and SAT-10 testing took place last for all
students, potentially opening results up to the influence of order effects. Second,
the participant pool was a convenient sample of assenting, high-performing
individuals and may not be representative of the larger public school population.
When compared to national samples, the group average for participants in this
sample was at the 63rd and 59th percentiles in SAT-10 science and social studies,
respectively. Third, with the sample consisting of students from a single grade,
caution is warranted when generalizing results to other grade levels.

**Implications**

With the present limitations and findings described, it is noteworthy
that evidence continues to support consideration of indicators of academic
language as viable GOM tools in the documentation of performance and
progress in science and social studies content. While still focused on Stage 1
(of 3) evaluation of the static score (Fuchs, 2004), the present findings lay a
foundation for a technically adequate, instructionally effective and logistically
feasible assessment framework in the content areas that include general
outcome measures of academic language and comprehension.

A first implication relates to the utility of academic vocabulary as a
measurement index in the content areas. A growing body of comparison
research in the GOM of content areas favor instruments like critical content
made descriptive comparisons of three predictors (i.e. vocabulary matching, maze, and oral reading fluency) and reported that the largest correlations with the criterion were primarily those with vocabulary matching. Mooney, McCarter, Schraven et al. (2013) made direct comparisons between the correlations of vocabulary matching, maze, and a writing GOM and found statistically significant differences in correlations with a statewide accountability test that favored vocabulary matching. Multiple regression analyses in these studies provided evidence that the measure of academic language provided the strongest explanatory power in the series of models analyzed. That pattern was repeated with critical content monitoring in the present study.

Furthermore, the utility of GOM with academic language as the index appears to extend to both science and social studies content areas. Initial research in this area focused on social studies learning, with linear relations with criterion measures across the upper elementary and middle school grades generally reported to be in the moderate to strong range (e.g., Espin et al., 2001). However, recent studies, including this one, have extended the evaluation of linear relationship to science across the elementary and secondary grades (e.g., Espin et al., 2013), with findings that have generally been moderate in magnitude and statistically significant, ranging from .45 to .66. The fact that a GOM index of content learning can be applied across subject areas bolsters the potential of the tool to be instructionally effective, something critical to RTI assessment frameworks. Moreover, logistical feasibility is apparent in the fact that the test can be administered using online technology. Such findings provide special educators with a research-informed choice when it comes to measuring academic performance in science and social studies classrooms.

A second implication relates to continued discovery of appropriate measurement technologies for use in upper elementary and secondary settings. The uses of GOM continue to grow across the K-12 spectrum. General outcome tools continue to be developed and evaluated as do assessment systems. Recently, Deno and colleagues (2009) described the implementation of a schoolwide elementary grades reading assessment program that incorporated universal screening and progress monitoring using maze and oral reading fluency measures in a manner that attempted to document progress across grades and over time. Similar efforts to develop long-term and sensitive assessment technologies in the content areas appear warranted.

Multiple measures appear worthy of continued exploration. Across science and social studies content, critical content monitoring demonstrated
the greatest explanatory power in regression models. Sentence verification technique, a measure of reading and possibly content comprehension, also was a significant predictor in both content areas. Commonality analyses (see Table 5) indicated that both measures had their own and shared substantial explanatory power for SAT-10 performance in science and social studies. Sentence verification technique has also proven influential in predicting variance in models of reading comprehension, along with measures such as maze, oral reading fluency, and written retell (Marcotte & Hintze, 2009). With a host of measures demonstrating the ability to explain variance in achievement, there is a need to explore the efficacy of various combinations of measures in terms of their collective technical adequacy, instructional effectiveness, and logistical feasibility (Deno & Fuchs, 1987). Such inquiry would build on as well as possibly inform the research being conducted in middle grades reading progress monitoring assessment (e.g., Barth et al., 2012; Tolar, Barth, Fletcher, Francis, & Vaughn, 2014) and likely impact both general and special education. Special educators can contribute to inquiry in classrooms across the continuum of alternative placements, particularly inclusive general education classroom settings.

References


Barth, A. E., Stuebing, K. K., Fletcher, J. M., Cirino, P. T., Francis, D. J., & Vaughn, S. (2012). Reliability and validity of the median score when assessing the oral reading fluency of middle grade readers. Reading Psychology, 33, 133-161.


**VALIDITY OF TWO GENERAL OUTCOME MEASURES OF SCIENCE AND SOCIAL STUDIES ACHIEVEMENT**

Paul Mooney, Renée E. Lastrapes, Amanda M. Marcotte, Amy Matthews, B. S.

**Summary**

Structured formative assessment in reading and mathematics has long been part of the fabric of special education. However, inquiry addressing the efficacy of structured formative assessment techniques to document student achievement and growth in social studies and science content is still in its infancy. Originally, reading measures were evaluated to determine their utility in measuring progress in the content areas. Over the past two decades the literature has expanded to include content focused instruments including vocabulary matching and content maze.

The present research addressed three more content-oriented instruments that have the advantage of being administered and scored online: Critical content monitoring, sentence verification technique, and written retell. The instruments were being evaluated for their technical adequacy and logistical
feasibility. Two research questions were evaluated: (a) what were the correlations with nationally-normed standardized achievement and statewide accountability tests in science and social studies content for critical content monitoring, sentence verification technique, and written retell? and (b) what was the incremental validity of adding measures of reading comprehension, using sentence verification technique and/or written retell, to content-focused achievement models that included critical content monitoring?

Participants were fifth-grade students (N = 51) enrolled in a public primary school in the southeastern U.S. As a whole, participants were 11.1 years old (SD = .5) at the time of testing, 68.6% female (n = 35), 66.7% Caucasian (n = 34), and 68.6% full-pay lunch status (n = 35). The three predictor variables were correlated with content test scores from the nationally representative standardized achievement test (i.e., Stanford Achievement Test-Tenth Edition abbreviated online form) and a statewide accountability test. Pearson correlations for critical content monitoring and the Stanford tests across science (r = .55) and social studies (r = .63) were moderately strong and similar in magnitude with other reported correlations for academic language measures in the literature. Correlations for critical content monitoring were descriptively larger than those between the standardized tests and sentence verification technique and written retell. Commonality analyses indicated that both critical content monitoring and sentence verification technique added unique variance to explanatory models.

A discussion of the results contributed to two implications. First, academic language, at the core of structured formative assessment instruments such as vocabulary matching and critical content monitoring, appears to be a viable avenue for continued inquiry. Second, given educational recommendations to rely on data from multiple sources in decision-making processes and present findings indicating that multiple instruments added unique variance to explanatory models, the use of multiple structured formative assessment measures in the development of content assessment frameworks appears warranted.

Study limitations included the order of presentation of the instruments, the small size and makeup of the sample, and the focus on one grade level.