THE STRUCTURE OF AN ONLINE ASSESSMENT OF SCIENCE AND SOCIAL STUDIES CONTENT: TESTING OPTIONAL FORMATS OF A GENERAL OUTCOME MEASURE

Paul Mooney, Kevin S. McCarter, Robert J. Russo, Danielle L. Blackwood
Louisiana State University, USA

Abstract

Research[1] questions addressed the structure of an online form of content-focused general outcome measurement known as critical content monitoring. The assessment tool is designed to serve as an accurate and efficient measure learning performance and progress in science and social studies classes. Scores from single administrations of critical content monitoring probes that varied in content and length were correlated with results from a statewide accountability content test for a sample of American fifth-grade students. The magnitude of correlation was moderate for probes that included a single content (i.e., social studies) or a mix of two subjects (i.e., social studies and science). Comparable correlations were reported for probes that featured reduced time or increased questions. Limitations and research implications are discussed.

Key Words: *Content Area Assessment, Formative Assessment, Technology*

The Structure of an Online Assessment of Science and Social Studies Content: Testing Optional Formats of a General Outcome Measure

Improving teaching practices as a means of increasing student outcomes for all and particularly struggling learners has been an international desire for decades. One area of educational stakeholder emphasis has been on the use of in-class assessments that are designed to inform teachers' next instructional steps. In his book *Embedded Formative Assessment*, Wiliam (2011) suggested that "attention to minute-by-minute and day-to-day formative assessment is likely to have the biggest impact on student outcomes" (p. 27). Positive and meaningful impact on the outcomes of students with or at risk for disabilities is critical to the future academic lives of this population given the academic delays that they face. The use of structured formative assessment has a strong evidence base to support its use by teachers and school systems (Stecker, Fuchs, & Fuchs, 2005). The present study focused on the efficacy of various forms of a structured formative assessment measure that are designed to measure performance and progress in the natural and social sciences. Specifically, the study addressed research questions related to different structural formats of critical content monitoring (Mooney, McCarter, Russo, & Blackwood, 2013) that vary according to the number of questions, length, and content make-up of a single probe.

Critical Content Monitoring

On the formative assessment continuum (Dorn, 2010), critical content monitoring fits on the formal or structured end as a frequent quantitative measure of specific skills. It is an

[1] The research reported here was supported by funding from the Louisiana Board of Regents, Louisiana Systemic Initiatives Program. The opinions expressed are those of the authors and do not represent views of the Louisiana Board of Regents

example of general outcome measurement (Fuchs & Deno, 1991) in that it targets an entire curriculum domain (e.g., 5th grade science) and develops equivalent tests that sample from the whole domain and indicate end-of-year skill or subject competence. In the measure's present form, upper elementary or middle school students read definitions of key grade-level science or social studies vocabulary at a secure learning management system link and are asked to choose the correct term name from a list of choices. Students are allowed up to 5 minutes to complete as many of the 20 questions as they are able. At the completion of the online quiz, students are provided their total score in number and visual (i.e., line graph) form. If a student has completed multiple probes, then the student views her or his scoring history.

One of eight curriculum-based measures described in the literature, critical content monitoring is an online adaptation of vocabulary matching (Espin & Deno, 1994-1995), a general outcome measure that has a body of technical adequacy findings supporting its use as a measure of content area performance and progress across grades (Espin, Busch, Lembke, Hampton, Seo, &Zukowski, 2013; Espin, Shin, & Busch, 2005; Mooney, McCarter, Schraven, &Callicoatte, 2013). One, moderate (.3-.7) to strong (.7-1.0; Reynolds, Livingston, & Wilson, 2009) correlations have been reported with standardized tests of achievement across multiple research teams (e.g., Espin,Busch, Shin, &Kruschwitz, 2001). Two, administrations of equivalent forms of the instrument over time have demonstrated that multiple versions are sensitive to student learning over time (e.g., Borsuk, 2010). Three, direct comparisons of vocabulary matching with other curriculum-based measures have demonstrated stronger magnitudes of correlations for vocabulary matching (Mooney, McCarter, Schraven, et al., 2013).

The promising efficacy findings relative to vocabulary matching are tempered by the reality that the general outcome measure of content learning still remains largely a researcher-driven technology two decades following its introduction. This stands in stark contrast to the wide-spread application of the one-minute oral reading fluency measure for screening and progress-monitoring purposes. A number of explanations have been posited for the lack of scaling of vocabulary matching. These reasons include the labor-intensiveness of vocabulary matching application as well as possible difficulties in interpreting some of the test score data (e.g., Mooney, McCarter, Schraven, et al., 2013).

One way to address potential barriers to broader use of content-focused screening and progress monitoring has been to incorporate technology into instructional practice. Possible applications include online progress-monitoring and/or instructional management systems and use of computer adaptive testing. In the area of curriculum-based assessment, computer technology has allowed for large-group administration of instruments, automatic test scoring, and real-time reporting of scores (Fuchs & Fuchs, 2001; Stecker et al., 2005). In the areas of math and science, the use of technology-enhanced formative evaluation systems have evidenced positive impacts on classroom grades and state accountability test results (Burns, Klingbeil, &Ysseldyke, 2010; Vannest, Parker, & Dyer, 2011; Ysseldyke & Bolt, 2007). Computer adaptive testing has been suggested as an effective tool for universal screening, with Ball and Christ (2012) describing research evidencing strong overall decision accuracy and individual skill information for computer adaptive reading tests.

Critical content monitoring is an online adaptation of vocabulary matching and conceptualized as a general outcome measure of content area learning. The target curriculum is an entire body or domain of academic language, with academic words assumed to be proxies for academic content learning. Academic language, defined as "the specialized language, both oral and written, of academic settings that facilitates communication and thinking about disciplinary content" (Nagy & Townsend, 2012, p. 92), is targeted for two reasons. First, as Nagy and Townsend (2012) point out, its proficiency allows students to better access meaning from academic text and discussion, achieve in school, and even act like scientists and historians. Alexander (n.d.) has equated vocabulary to communicative currency. Second, from an intervention perspective, academic language is an alterable variable (Bloom, 1980) that is particularly pertinent to struggling learners, who Baker, Kame'enui, and Simonsen (2007) indicate generally know fewer words at instruction's inception and learn fewer words over the course of instruction in comparison to regularly achieving peers. Academic vocabulary development provides struggling learners and those who teach them the chance to both build background knowledge and promote active engagement in the learning environment. The instructional purposes, materials, and practices that derive from academic vocabulary application also lend themselves to meaningful differentiation for the diverse populations that comprise secondary school learning environments.

Mooney, McCarter, Russo, et al. (2013) reported findings on the initial application of critical content monitoring, with research addressing two primary technical adequacy questions related to the static score. The first query related to the magnitude of the correlations between scores from the online assessment and a statewide accountability subject test. Twenty separate online science probes were completed in close proximity to each other by a convenience sample of generally high-performing students, with scores then correlated with the standard score of the statewide science accountability test that was administered the following week. Results indicated moderate correlations for the 20 probes, with single correlations ranging from .36 to .55 and a pooled estimate for participants that completed all 20 probes determined to be .45 (Mooney, McCarter, Russo, et al., 2013). Correlation magnitudes between the online science probe and state accountability test were lower than those reported in the vocabulary matching literature across both science and social studies content (e.g., Espin et al., 2013).

The second research question addressed the equivalence of the probes as developed. Twenty probes were administered over a 10-day time frame in order to determine whether there were any differences in the scores of probes that were created using identical processes. The equivalence of probe scores is believed to be a critical feature of general outcome measures designated as indexes of growth (Petscher, Cummings, Biancarosa, & Fien, 2013). If test scores from individual probes are not determined to be equivalent, then it may be difficult to ascertain if changes in scores are the result of learning alone or some combination of variables which may or may not include student learning. Results of comparisons of the 20 online probes indicated that while descriptive differences in the correlations proved statistically equivalent, there were statistically significant differences across the online probe's mean scores (Mooney, McCarter, Russo, et al., 2013). In fact, the pattern of scores indicated increases over the 10 days of testing, suggesting evidence of student learning of some form. Findings of non-equivalence in scoring for probes previously assumed to be parallelwere similar to results reported in the general outcome measurement literature related to reading-CBM (e.g., Ardoin, Williams, Christ, Klubnik, & Wellborn, 2010).

Object of the research – to reveal the structure of an online form of content-focused general outcome measurement as critical content monitoring.

Specifically, the study addressed research questions related to different structural formats of critical content monitoring (Mooney, McCarter, Russo, & Blackwood, 2013) that vary according to the number of questions, length, and content make-up of a single probe.

Research Rationale and Questions

To determine whether data from curriculum-based measures of student performance effectively and meaningfully impact data-based decision making, Deno and Fuchs (1987) proposed a framework that encompassed a series of technical adequacy, instructional effectiveness, and logistical feasibility questions targeting issues of what to measure, how to measure, and how to use data. The initial critical content monitoring findings fell largely in the technical adequacy section of the Deno and Fuchs (1987) matrix, with their focus on validity and reliability issues for single comparisons. These initial findings generated a number of questions that prompted some logistical feasibility questions that provided a rationale for the present inquiry.

In describing logistical feasibility, Deno and Fuchs (1987) noted that measurement systems are best served when they reduce teacher and student time in measurement. Of primary concern in the present inquiry was the structure of the probe within the context of validity research. The first question related to subject matter. Previously, science vocabulary populated the probe and was evaluated for its association with a state accountability test standard score. The present research evaluated a social studies-oriented probe and accompanying predictor-criterion relationship. A second question related to subject matter mix. In the content area learning literature, research has focused on single grade-level subject areas. As a result, science and social studies content are assessed separately and the time associated with content assessment then doubled when both subjects are included in assessment systems. One result of an increase in testing time is an accompanying reduction in time allocated to and available for instruction. The present study evaluated the criterion validity of a single probe that included both science and social studies vocabulary within the traditional 5-minute time frame.

A third question also addressed assessment efficiency and derived from a qualitative review of participant performance in Mooney, McCarter, Russo,et al. (2013). That is, while participants were allowed up to 5 minutes to complete the science focused online probes, they uniformly completed the process in less than 3 minutes. In an effort to determine whether the length of administration time could be varied without impacting the relationship to a meaningful criterion, researchers evaluated the administration of probes that differed in terms of time allowed. A final structure-oriented query addressed the number of questions in a probe. In response to findings indicating weaker criterion validity relationships for critical content monitoring probes in comparison to vocabulary matching measures, the impact of doubling the number of questions was evaluated in terms of its relationship to a state accountability test result. Salvia, Ysseldyke, and Bolt (2007) indicated that test reliability interpretations generally are strengthened by including more questions.

Method

Participants and Setting

The Institutional Review Board-approved study involved the same population of students in Mooney, McCarter, Russo, et al. (2013). Participants were fifth-graders (N = 106) in a single public K-12 school in south Louisiana. As a whole, participants were 58 percent female and 88 percent Caucasian. All participants paid full price for school lunches. No students were verified with exceptionalities. As a group, the students demonstrated success in school, with the median and most commonly reported quarter grade in the A- range. All but one of the participants (99%) scored at the basic level of proficiency on both science and social studies tests. Participants were taught in a departmentalized setting. Testing took place during science class.

As detailed in Mooney, McCarter, Russo et al. (2013), the sample was chosen because of the timing and time-intensive nature of the initial pilot research. That is, approximately 15 minutes of instructional time was utilized daily over 10 successive days in the two weeks before statewide accountability testing in order to collect the data necessary to answer the study's proposed research questions. It was accepted by the researchers that the possible negative impact of taking instructional time away from students and the teacher for this generally high-achieving sample outweighed legitimate concerns related to the generalizability of results to other public-school student populations. With the same sample of students having taken part in the present research, caution, then, is warranted in evaluating the study's results.

Measures

Two measures were compared in the present study: (a) different forms of critical content monitoring for fifth-grade content; and (b) subject tests of the criterion-referenced *integrated* Louisiana Educational Assessment Program (*i*LEAP; LDE, n.d., a).

Critical content monitoring. The content-focused general outcome measure described earlier evolved from procedures previously outlined in Espin et al.(2001). It is one of two online content-focused assessment systems that have been described in the literature, along with key vocabulary progress monitoring (Vannest et al., 2011). Critical content monitoring probes were developed using a curriculum sampling approach (Fuchs, 2004). Terms in each probe were randomly selected from the full body of content terms by unit, with each probe including terms from each unit. 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. Adjustments were made to the number of terms per unit if the total number of terms for all of the units did not add up to the probe question total. For the present study, assessment formats differed in the number of questions included per probe, the course content in each probe, and the length of administration. Terms and accompanying definitions were entered into an online learning management system (Moodle, n.d.) in a multiple-choice format. The Moodle system utilized in this project was extensively customized to evaluate student performance in real time and show trends in student performance graphically.

i*LEAPgrade 5 criterion-referenced test.* The stated purpose of *i*LEAP is measurement of students' proficiency in reaching Louisiana academic standards in English/language arts, math, science, and social studies (LDE, n.d., a). It has been Louisiana's statewide assessment for 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 science, life science, earth and space science, and science and the environment, with test questions addressing all five 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. Students scoring at or above the basic level were considered to have passed the test. Technical adequacy data for the *i*LEAP fifth grade tests were accessed from the LDE website. Cronbach's alpha levels of 0.85 for science and 0.82 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 (LDE, n.d., b). The correlation between the fifth-grade *i*LEAP science test and the science test from the abbreviated online Stanford Achievement Test Series (10th ed., Pearson Education, n.d.) was .64 (Mooney, 2014).

Procedure

Participants were administered a number of different critical content monitoring forms over a five-day period in mid-May 2011, near the end of the school year and about six weeks after statewide accountability testing. Probes differed in content, number of questions, and/or length of administration. Testing followed procedures similar to those described in Mooney, McCarter, Schraven et al. (2013). That is, the order of presentation of probes for each of 4 fifth-grade sections was chosen through random selection without replacement in order to address possible order effects. Students logged in to the secure Moodle site using an individual login and password. The teacher verbally guided students to the site first and then the actual probe. After delivering a standard instruction, students were expected to access the probe and answer the questions within the time limit.

Probe scores were accessed from the online site by the first author and exported in the form of individual Excel files once the testing process was complete. One Excel file included all final scores for participants and was used to complete criterion-related validity analyses. Student demographic and statewide accountability test data were the same as those summarized in Mooney, McCarter, Russo,et al. (2013). Because the probes were computer scored, no inter scorer reliability actions were formally taken for the online test scores beyond a double checking of the entry of the questions into the online system to ensure that the right choice accompanied each stem.

Data Analysis

Across the four questions, correlation analysis was used to quantify the linear relationship between the various critical content monitoring forms and the statewide content tests. Point estimates and 95% confidence interval (CI) estimates of the true, unknown correlations were computed from the data. Confidence intervals were used to test whether a significant linear relationship existed between variables, such that if the respective CI estimate did not include zero then it was concluded that there existed a linear relationship.

Results

Question 1: Social Studies

Table 1 provides means, SDs, correlations, and 95% CI estimates concerning critical content monitoring probes with social studies content. In descriptive terms, the correlation between scores of the traditional (i.e., 5 minute) critical content monitoring and *i*LEAPsocial studies test was moderate in magnitude [i.e., .67 (95% CI .55, .77)].

Question 2: Mixed Content

Tables 1 and 2 provide means, SDs, correlations, and 95% CI estimates for two forms of critical content monitoring probes with mixed content. One form was 5 minutes in length and comprised of 20 questions that included half-social studies vocabulary and half-science terms that were introduced in alternating fashion. A second form used the 50/50 content split but was 3 minutes and 40 questions in length. For social studies and science, correlations associated with scores from the 5-minute probe were descriptively greater in magnitude than those from the 3-minute, 40-question probe.

Question 3: Reduced Time

Table 2 provides the mean, SD, correlation, and 95% CI estimate for a 3-minute critical content monitoring probe of science content. The correlation with *i*LEAP science was moderate in magnitude and descriptively comparable to the probe-*i*LEAP correlation for the traditional 5-minute critical content monitoring probe.

Question 4: Increased Questions

Table 2 also provides the mean, SD, correlation, and 95% CI estimate for a 40-question critical content monitoring probe of science content. The correlation with *i*LEAP science was moderate in magnitude and descriptively smaller than the probe-*i*LEAP correlation for the traditional 5-minute critical content monitoring probe.

Discussion

In extending initial critical content monitoring validity research, the present findings set the stage for some interesting implementation possibilities moving forward. Answers to the four research questions will be summarized in relationship to the larger content area general outcome measurement literature prior to a discussion of limitations and implications.

All of the study's research questions focused, in some manner, on the structure of the critical content monitoring probe. The first question addressed outcomes of a change in content focus of a 5-minute probe from science to social studies, with results indicating that the magnitude of the linear relationship with a state content test criterion increased in that circumstance. The correlation of .67 (95% CI .55, .77) was moderate in magnitude (Reynolds et al., 2009) and descriptively larger than the .54 correlation for science reported in the present study (see Table 2) and the range of correlations (i.e., .36 to .55; pooled mean = .45) reported in Mooney, McCarter, Russo, et al. (2013). Moreover, the linear relationship between the social studies probe and test scores was comparable to correlations with meaningful criterion measures (e.g., Espin et al., 2001) reported in the vocabulary matching literature.

The second question also addressed social studies content but within the context of a mix of social studies and science content within a single probe. Correlations with a state test for a 5-minute mixed probe were descriptively comparable (i.e., *r* = .66; 95% CI .53, .76) to those of a strict social studies content measure and, again, similar to vocabulary matching magnitudes in the literature. The pattern of descriptively larger correlations for social studies over science content was also evident for the 5-minute, mixed-content probe. However, that pattern was not maintained for the 3-minute, 40-question mixed-content probe.

The third question addressed a probe adjustment in which the time to complete the probe was decreased. Correlations between 3- and 5-minute probes and state tests in science content were generally comparable, with similar confidence intervals and mean scores as well. However, there were relatively large descriptive declines in the magnitude of correlations for reduced-time probes that also included mixed content and an increased number of questions (see Tables 1 and 2).

Table 1. Critical Content Monitoring (CCM) Means [with Standard Deviations (SD)], and Correlations* with State Social Studies Accountability Test [each with 95% Confidence Interval (CI)] in Fifth Grade

Probe	N	Mean	SD	95% CI	r	95% CI
CCM Traditional: Social Studies	102	14.3	3.1	[13.7, 14.9]	.67	[.55, .77]
CCM ½ Science/½ Social Studies Mix	100	16.3	2.8	[15.7, 16.8]	.66	[.53, .76]
CCM Science/Social Studies-40 Questions-3 Minutes	102	20.5	7.4	[19.0, 21.9]	.36	[.18, .52]

* *p*< .01.

Table 2. Critical Content Monitoring (CCM) Means [with Standard Deviations (SD)], and Correlations* with State Science Accountability Test [each with 95% Confidence Interval (CI)] in Fifth Grade

Probe	N	Mean	SD	95% CI	*r*	95% CI
CCM Traditional: Science	99	18.2	1.9	[17.8, 18.6]	.54	[.39, .67]
CCM Science-3 Minutes	101	17.8	2.5	[17.4, 18.3]	.53	[.37, .65]
CCM Science-40 Questions	100	29.6	4.4	[28.7, 30.4]	.47	[.30, .61]
CCM ½ Science/½ Social Studies Mix	100	16.3	2.8	[15.7, 16.8]	.62	[.48, .72]
CCM Science/Social Studies-40 Questions-3 Minutes	102	20.5	7.4	[19.0, 21.9]	.37	[.19, .53]

* *p*< .01.

For science and social studies content, correlational magnitudes were on the low – as opposed to the high – end of the moderate range. The final question addressed the impact of increasing the number of questions for a critical content monitoring probe. Findings for a 40-question science content probe indicated that the correlation magnitude was descriptively smaller than the correlation for the 20-question test. However, it did appear as though students were able to correctly answer more questions in the 5-minute time frame when faced with 40 questions versus 20 questions for the traditional science probe. Interestingly, when the content was mixed and the time-to-complete reduced, there were also a larger number of questions correctly answered than for a reduced-time science-only probe.

Limitations

All present findings must be interpreted in a context in which participants were generally high-achieving and not likely representative of most public school settings. That noted, there were legitimate reasons for seeking out this population of students and conducting the type of research described herein. A reader's confidence in the reported findings will no doubt be strengthened if there are comparable results given more diverse samples, particularly with populations including students with and/or at risk for high-incidence disabilities. Generalizability concerns have been commonly reported in the general outcome measurement literature for content area learning given that the bulk of research to date has involved single subjects and grade levels and predominantly been directed by researchers. A final limitation related to the timing of testing. Namely, the magnitude of correlations with *i*LEAP reported in the present study may have been impacted by the additional instruction that took place between the statewide testing and the second administration of critical content monitoring tests. Furthermore, with the testing taking place at the end of the school year, concerns could be raised about the effort of the students. However, correlations with the state test for an identical critical content monitoring science probe that was administered at state test time and again at the end of the school year were comparable, suggesting that similar effort was offered by participants.

Implications

While recognizing the identified generalizability concerns, researchers believe there are some implications emanating from the present findings and those in the content-focused general outcome measurement literature that are worthy of consideration for all students, including those with or at risk for disabilities. First, there continues to be the possibility that online assessment systems can be adapted for use in content classrooms and as general outcome measures of content learning. The present findings and those of Mooney, McCarter, Russo, et al. (2013) and Vannest et al. (2011) suggest that online technologies can be used to reliably and validly approximate student content learning performance and progress. That is, generally moderate correlations were reported for meaningful measures of performance in social studies and science – full range .25 to .83 – across studies; there was evidence of variable growth for participants across time (Vannest et al., 2011); and users reported successfully navigating and liking an online system (Mooney, McCarter, Russo, et al., 2013). Taken together with the successful applications of formative evaluation and computer adaptive systems in science, math, and reading (e.g., Ball & Christ, 2012; Burns et al., 2010), it appears that comprehensive, efficient, and effective screening and progress monitoring systems across secondary school settings are both possible and worthy of researcher, practitioner, and/or commercial entity resource allocation.

Second, it looks like alterations to the structure of online general outcome measures merit continued inquiry on empirical and practical grounds. Results of the present static score research indicate that it may be feasible to administer probes that mix science and social studies content over 5 minutes or reduce to 3 minutes the time for a single subject probe. As a consequence, then, there can be greater coverage of content (i.e., two subjects versus one), more curriculum-based data for decision making of a potentially formative nature, and more allocated time for instruction that could be adjusted given the real-time access to data that online systems provide. While the focus of content area instruction should not be isolated to vocabulary teaching – and is not advocated through presentation of these results – there remains the opportunity to differentiate instruction based on the results of general outcome measures of content learning that incorporate academic vocabulary as proxies for learning. A strong body of research in vocabulary instruction incorporating explicit instruction, resource access, and immersion in rich environments exists to improve the achievement of struggling learners, including students with or at risk for high-incidence disabilities (Carlisle, Kenney, & Vereb, 2013).

Third, and related to progress monitoring, there may be reason to continue exploration of the increased number of question alteration in critical content monitoring probes. While there appeared to be a descriptive detrimental impact of doubling the number of questions to the probe on the static score correlation with a criterion, the fact that there were increased mean scores for both lengthened probes may prove beneficial in terms of future growth-oriented research. That is, larger final or subsequent probe mean scores for populations of students, when compared to beginning or earlier probe mean scores, could mean that the average weekly growth rates better approximate meaningful information to teachers and students than some of the rates previously reported in the vocabulary matching literature (e.g., Beyers, Lembke, & Curs, 2013; Mooney, McCarter, Schraven et al., 2013). With a 40-question probe and 36 weeks in a school year, it is at least possible to generate a 1-word-per-week growth rate. Teachers, students, parents, paraprofessionals, and administrators would likely better understand, appreciate, and believe that they could manipulate through goal setting and intervention a growth rate for a tangible 'product' (i.e., academic vocabulary) that was .5 or higher (as has been reported by Espin et al., 2005; 2013) than rates of .26 and lower (as reported in Beyers et al., 2013; Borsuk, 2010; and Mooney, McCarter, Schraven et al., 2013). With larger growth rates, learning as indicated in increasing general outcome measure scores could be observed in 1-2 week increments as opposed to 4-10 week time frames, making the data more meaningful to consumers. The utility of increasing the number of questions within the 5-minute time frame is deserving of continued inquiry.

Finally, and related to the previous two points on a larger scale, the online application of general outcome measurement using academic vocabulary in content classrooms provides interested stakeholders the chanceto contribute to the development of a broad-based manageable and meaningful framework for documenting achievement in secondary schools. At its core, though developed for purposes of screening and progress monitoring and described as a proxy for learning in the tradition of general outcome measurement, critical content monitoring is an assessment of academic words. Nagy and Townsend (2012) describe words as tools for learning; Alexander (n.d.) refers to them more broadly as the currency of communication. And words, in multiple forms, are content in all secondary courses. While the assessment of words (or academic vocabulary in this case) has a storied history, Pearson, Hiebert, & Kamil (2007) argue that its research has been "grossly undernourished, both in its theoretical and practical aspects" (p. 282).

Vocabulary assessments can be categorized along three continua introduced by Read (2000): (a) in terms of construct, from discrete or by itself as a construct (e.g., vocabulary knowledge) to embedded within a larger construct (e.g., vocabulary's contribution to comprehension); (b) in terms of the nature of what is to be learned, selective (or isolated) to comprehensive (or all-encompassing); and (c) in terms of the context of the question, context-dependent (or needing to use the context) to context-independent (or not needing context).

The authors contend that, as a general outcome measure of content learning, critical content monitoring is a vocabulary assessment measure that falls on the embedded, comprehensive, and context-independent ends of the three continua. That combination of design elements, including a proportional sampling from an entire content curriculum and the conceptualization of the test score as a vital sign of academic learning (Deno, 1985), may be the reason why a listing of words and definitions evidenced reasonably strong initial linear relationships to meaningful criteria, though limited to a single grade level, statewide test, and high-achieving sample.

Conclusion

A great deal more focused inquiry derived from the principles of general outcome measurement remains warranted and seems possible given the online capabilities that have been validated, to some degree, in the content area assessment literature. Online technologies will allow for the evaluation of different forms of question, for example, using stems that could target definitional or application or evaluative language. Online technologies might manipulate context-dependent versus context-independent delivery approaches. Online technologies might even investigate grade-level versus multi-grade-level content probes in an effort to make the system more efficient and less segregated. Research in this context may allow what has, to this point in time, been special education-driven inquiry, influenced by the framework of Deno and Fuchs (1987), to reach outside the general outcome measurement walls to inform and be further shaped by the larger theoretical and practical work taking place in vocabulary assessment and instruction.

References

1. Alexander, F. (n.d.). Understanding vocabulary. Retrieved from: http://www.scholastic.com/teachers/article/understanding-vocabulary.
2. Ardoin, S. P., Williams, J. C., Christ, T. J., Klubnik, C., & Wellborn, C. (2010). Examining readability estimates' predictions of students' oral reading rate: Spache, lexile, and forcast. *School Psychology Review*, *39*, 277-285.
3. Baker, S. K., Kame'enui, E. J., Simmons, D. C., &Simonsen, B. (2007). Characteristics of students with diverse learning and curricular needs. In M. D. Coyne, E. J. Kame'enui, & D. W. Carnine (Eds.), *Effective teaching strategies that accommodate diverse learners* (3rded.)(pp. 23-43). Upper Saddle River, NJ: Pearson.
4. Beyers, S. J., Lembke, E. S., & Curs, B. (2013). Social studies progress monitoring and intervention for middle school students. *Assessment forEffective Intervention*, *38*, 224-235. doi: 10.1177/1534508413489162
5. Borsuk, E. R. (2010). Examination of an administrator-read vocabulary-matching measure as an indicator of science achievement. *Assessment for Effective Intervention*, *35*, 168-177. doi: 10.1177/1534508410372081
6. Burns, M. K., Klingbeil, D. A., & Ysseldyke, J. (2010). The effects of technology-enhanced formative evaluation on student performance on state accountability math tests. *Psychology in the Schools*, *47*(6), 582-591.doi: 10.1002/pits
7. Carlisle, J. F., Kenney, C. K., &Vereb, A. (2013). Vocabulary instruction for students at risk for reading disabilities: Promising approaches for learning words from texts. In D. J. Chard, B. G. Cook, & M. Tankersley (Eds.), *Research-based strategies for improving outcomes academics* (pp. 44-57). Boston: Pearson.
8. Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*, 219-232.
9. Deno, S. L., & Fuchs, L. S. (1987). Developing curriculum-based measurement systems for data-based special education problem solving. *Focus on Exceptional Children*, *19*(8), 1-16.
10. Dorn, S. (2010).The political dilemmas of formative assessment.*Exceptional Children*, *76*, 325-337.

11. Espin, C. A., Busch, T. W.,Lembke, E. S., Hampton, D. D., Seo, K., &Zukowski, B. A. (2013). Curriculum-based measurement in science learning: Vocabulary-matching as an indicator of performance and progress. *Assessment for Effective Intervention*, *38*, 203- 213. doi: 10.1177/1534508413489724
12. Espin, C. A., Busch, T. W., Shin, J., & Kruschwitz, R. (2001). Curriculum-based measurement inthe content areas: Validity of vocabulary matching as an indicator of performance in social studies. *Learning Disabilities Research & Practice*, *16*, 142-151. doi: 10.1111/0938-8982.00015
13. Espin, C. A., & Deno, S. L. (1994-1995). Curriculum-based measures for secondary students: Utility and task specificity of text-based reading and vocabulary measures for predicting performance on content-area tasks. *Diagnostique*, *20*, 121-142.
14. Espin, C. A., Shin, J., & Busch, T. W. (2005). Curriculum-based measurement in the content areas: Vocabulary matching as an indicator of progress in social studies learning. *Journal of Learning Disabilities*, *38*, 353-363.
15. Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*, 188-192.
16. Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, *57*, 488-500.
17. Fuchs, L. S., & Fuchs, D. (2001). Computer applications to curriculum-based measurement. *Special Services in the Schools*, *17*(1-2), 1-14.doi: 10.1300/J008v17n01_01
18. Ketterlin-Geller, L. R., McCoy, J. D., Twyman, T., & Tindal, G. (2006).Using a concept maze to assess student understanding of secondary-level content. *Assessment for Effective Intervention*, *31*(2), 39-50.doi: 10.1177/073724770603100204
19. Klingner, J. K., Vaughn, S., Dimino, J., Schumm, J. S., & Bryant, D. P. (2001). *Collaborative strategic reading: Strategies for improving comprehension*. Longmont, CO: Sopris West.
20. Louisiana Department of Education.(LDE; n.d., a). Integrated Louisiana Educational Assessment Program (*i*LEAP). Retrieved from: http://www.louisianaschools.net/lde/uploads/9725.pdf
21. Louisiana Department of Education.(LDE; n.d., b).*iLEAP2010 Technical Summary.* Retrieved from: http://www.louisianaschools.net/lde/uploads/18005.pdf
22. Moodle (n.d.). Available: http://docs.moodle.org/23/en/About_Moodle
23. Mooney, P. (2014). Unpublished raw data.
24. Mooney, P., Benner, G. J., Nelson, J. R., Lane, K. L., & Beckers, G. (2008). Standard protocol and individualized remedial reading interventions for secondary students with emotional and behavioral disorders. *Beyond Behavior*, *17*(2), 3-10.
25. Mooney, P., McCarter, K. S., Russo, R. J., & Blackwood, D. L. (2013). Examining an online content general outcome measure: Technical features of the static score. *Assessment for Effective Intervention*, *38*, 249-260.doi: 10.1177/1534508413488794
26. Mooney, P., McCarter, K. S., Schraven, J., & Callicoatte, S. (2013). Additional performance and progress validity findings targeting the content-focused vocabulary matching. *Exceptional Children*, *80* (1), 85-100.
27. Mooney, P., McCarter, K. S., Schraven, J., &Haydel, B. (2010). The relationship between content area GOM and statewide testing in world history. *Assessment for Effective Intervention*, *35*, 148-158.doi: 10.1177/1534508409346052
28. Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47* (1), 91-108.doi: 10.1002/RRQ.011
29. Parker, R. I., Vannest, K. I., Davis, J. L., & Clemens, N. H. (2012). Defensible progress monitoring data for medium- and high-stakes decisions. *The Journal of Special Education*, *46*, 141-151. doi: 10.1177/0022466910376837
30. Pearson Education (n.d.). Stanford Achievement Test Series, online abbreviated form (10th ed.). Retrieved from: http://www.pearsonassessments.com/learningassessments/products/100000563/stanford-achievement-test-series-tenth-edition-abbreviated-battery.html
31. Pearson, P. D., Hiebert, E. H., &Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, *42*, 282-296. doi: 10.1598/RRQ.42.2.4
32. Petscher, Y., Cummings, K. D., Biancarosa, G., &Fien, H. (2013). Advanced (measurement) applications of curriculum-based measurement in reading. *Assessment for Effective Intervention*, *38*, 71-75.doi: 10.1177/1534508412461434

33. Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: The Cambridge Press.
34. Salvia, J., Ysseldyke, J. E., & Bolt, S. (2007). *Assessment in special and inclusive education* (10th ed.). Boston: Houghton Mifflin.
35. Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*, 795- 819. doi: 10.1002/pits.20113
36. Tindal, G. A., & Germann, G. (1991). Mainstream consultation agreements in secondary schools. In G. Stoner, M. R. Shinn, & H. M. Walker (Eds.), *Interventions for achievement and behavior problems* (pp. 495-518). Bethesda, MD: National Association of School Psychologists.
37. Vannest, K. J., Parker, R., & Dyer, N. (2011). Progress monitoring in Grade 5 science for low achievers. *The Journal of Special Education*, *44*, 221-233.
38. Wiliam, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.
39. Ysseldyke, J., & Bolt, D. M. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review*, *36*, 453-467.

THE STRUCTURE OF AN ONLINE ASSESSMENT OF SCIENCE AND SOCIAL STUDIES CONTENT: TESTING OPTIONAL FORMATS OF A GENERAL OUTCOME MEASURE

Summary

Paul Mooney, Kevin S. McCarter, Robert J. Russo, Danielle L. Blackwood
Louisiana State University, USA

Improving teaching practices as a means of increasing student outcomes for all and particularly strugglinglearners has been an international desire for decades. One area of educational stakeholder emphasis has been on the use of in-class assessments that are designed to inform teachers' next instructional steps. In his book *Embedded Formative Assessment*, Wiliam (2011) suggested that "attention to minute-by-minute and day-to-day formative assessment is likely to have the biggest impact on student outcomes" (p. 27). Positive and meaningful impact on the outcomes of students with or at risk for disabilities is critical to the future academic lives of this population given the academic delays that they face. The use of structured formative assessment has a strong evidence base to support its use by teachers and school systems (Stecker, Fuchs, & Fuchs, 2005). The present study focused on the efficacy of various forms of a structured formative assessment measure that are designed to measure performance and progress in the natural and social sciences. **Object of the research –** to reveal the structure of an online form of content-focused general outcome measurement as critical content monitoring.

Specifically, the study addressed research questions related to different structural formats of critical content monitoring (Mooney, McCarter, Russo, & Blackwood, 2013) that vary according to the number of questions, length, and content make-up of a single probe. On the formative assessment continuum (Dorn, 2010), critical content monitoring fits on the formal or structured end as a frequent quantitative measure of specific skills. It is an example of general outcome measurement (Fuchs &Deno, 1991) in that it targets an entire curriculum domain (e.g., 5th grade science) and develops equivalent tests that sample from the whole domain and indicate end-of-year skill or subject competence. One way to address potential barriers to broader use of content-focused screening and progress monitoring has been to incorporate technology into instructional practice. Possible applications include online progress-monitoring and/or instructional management systems and use of computer adaptive testing. In the area of curriculum-based assessment, computer technology has allowed for large-group administration of instruments, automatic test scoring, and real-time reporting of scores (Fuchs & Fuchs, 2001; Stecker et al., 2005).

Critical content monitoring is an online adaptation of vocabulary matching and conceptualized as a general outcome measure of content area learning. The target curriculum is an entire body or domain of academic language, with academic words assumed to be proxies for academic content learning. Academic language, defined as "the specialized language, both oral and written, of academic settings that facilitates communication and thinking about disciplinary content" (Nagy & Townsend, 2012, p. 92), is targeted for two reasons. First, as Nagy and Townsend (2012) point out, its proficiency allows students to better access meaning from academic text and discussion, achieve in school, and even act like scientists and historians. Alexander (n.d.) has equated vocabulary to communicative currency. Second, from an intervention perspective, academic language is an alterable variable (Bloom, 1980) that is particularly pertinent to struggling learners, who Baker, Kame'enui, and Simonsen (2007) indicate generally know fewer words at instruction's inception and learn fewer words over the course of instruction in comparison to regularly achieving peers.

The second research question addressed the equivalence of the probes as developed. Twenty probes were administered over a 10-day time frame in order to determine whether there were any differences in the scores of probes that were created using identical processes. The equivalence of probe scores is believed to be a critical feature of general outcome measures designated as indexes of growth (Petscher, Cummings, Biancarosa, &Fien, 2013).

Participants were fifth-graders (N = 106) in a single public K-12 school in south Louisiana. As a whole, participants were 58 percent female and 88 percent Caucasian. All participants paid full price for school lunches. No students were verified with exceptionalities. As a group, the students demonstrated success in school, with the median and most commonly reported quarter grade in the A- range. All but one of the participants (99%) scored at the basic level of proficiency on both science and social studies tests. Two measures were compared in the present study: (a) different forms of critical content monitoring for fifth-grade content;and (b) subject tests of the criterion-referenced *integrated* Louisiana Educational Assessment Program (*i*LEAP; LDE, n.d., a). Research let to draw some conclusions: A great deal more focused inquiry derived from the principles of general outcome measurement remains warranted and seems possible given the online capabilities that have been validated, to some degree, in the content area assessment literature. Online technologies will allow for the evaluation of different forms of question, for example, using stems that could target definitional or application or evaluative language. Online technologies might manipulate context-dependent versus context-independent delivery approaches. Online technologies might even investigate grade-level versus multi-grade-level content probes in an effort to make the system more efficient and less segregated. Research in this context may allow what has, to this point in time, been special education-driven inquiry, influenced by the framework of Deno and Fuchs (1987), to reach outside the general outcome measurement walls to inform and be further shaped by the larger theoretical and practical work taking place in vocabulary assessment and instruction.