Investigating Sentence Verification Technique as a Potential Curriculum-Based Measure of Science Content

Renée E. Lastrapes and Paul Mooney

Abstract: *This study examined the viability of the Sentence Verification Technique (SVT; Royer, Hastings, & Hook, 1979) assessment tool as a curriculum-based measure of science content learning. Its perceived compatibility with statewide accountability test expectations made SVT a candidate for use in content-focused response to intervention frameworks. Each SVT probe included two science-focused reading passages and 32 accompanying questions (16 per passage). Participants included 486 students in Grades 4 through 6 in a rural school district. Concurrent criterion validity correlations with one state accountability and one nationally standardized test of science content across three grades ranged from .18 to .55, with predictive correlations ranging from -.09 to .57. Internal consistency reliability correlations for the full probe ranged from .43 to .83. Test-retest reliability estimates across three grades ranged from .68 to .70. Study results inform discussions about the nature and structure of secondary school response to intervention assessment frameworks.*

According to the Nation's Report Card (2017), only 36% of public school students performed at or above proficiency level in Grade 4 reading assessments and 34% in Grade 8. These low proficiency statistics are problematic because, as students progress through primary education, there is an increased expectation that they learn content through reading (Espin et al., 2013). A potential mismatch exists between the reading proficiency level of many students and the difficulty level of the textbooks used as learning materials (Espin et al., 2013). Textbooks used in content classes such as science and social studies, in particular, often contain more advanced vocabulary and introduce a wide range of topics within a short amount of written space. Espin et al. (2013) suggested that these tools are "inconsiderate" of student and teacher needs.

Systematic work to design and evaluate interventions to address the difficulties of at-risk learners are ongoing (e.g., Boudreaux-Johnson, Mooney, & Lastrapes, 2017; Fuchs & Fuchs, 2019; Vaughn, Roberts, Miciak, Taylor, & Fletcher, 2019; Wexler, Reed, Barton, Mitchell, & Clancy, 2018). Separate from the interventions themselves, efforts exist to design and validate structured formative assessment instruments for use at the student, class, and school levels as part of upper elementary and secondary response to intervention (RTI) frameworks and alongside existing accountability structures (e.g., Barth, Tolar, Fletcher, & Francis, 2014; Conoyer et al., 2018; Johnson, Semmelroth, Allison, & Fritsch, 2013; Mooney & Lastrapes, 2018b). Given the reported reading skill delays of the majority of public school students, teachers need formative assessment instruments across content areas to help them make more timely decisions about appropriate instructional programming.

Curriculum-based measurement (CBM; Deno, 1985; Hosp, Hosp, & Howell, 2016) is an instructional assessment framework that was designed to document academic learning (Ford, Conoyer, Lembke, Smith, & Hosp, 2018; Mooney & Lastrapes, 2018a, 2018b). Curriculum-based measurement is a type of formative assessment that targets an entire school year curriculum and evaluates what a student should know or demonstrate by the end of a grade or subject. As part of the framework, purportedly equivalent measures, known as probes, are administered periodically to determine end-of-year competence in a subject area (Fuchs & Deno, 1991; Mooney & Lastrapes, 2018a, 2018b; Mooney, McCarter, Schraven, & Callicoatte, 2013). Efforts to examine different assessments to determine whether they will function as a CBM for specific content areas have increased, particularly related to science curricula (Borsuk, 2010; Espin et al., 2013; Ford et al., 2018; Ford & Hosp, 2017; Mooney & Lastrapes, 2016; Mooney, Lastrapes, Marcotte, & Matthews, 2016).

Sentence Verification Technique (SVT; Royer, Hastings, & Hook, 1979) has long been utilized as an assessment of reading comprehension (e.g., Marcotte, Rick, & Wells, 2019; Royer, 2005). The measure is based on a constructivist theory of reading comprehension, emphasizing the interaction between an incoming linguistic message and a reader's world knowledge, where the meaning of a message could be maintained in memory even if the exact structure of the message was not (Ford & Hosp, 2017; Marcotte et al., 2019). Reliability of the SVT has been investigated with Cronbach's alpha measures ranging from .5 to .9, with reliability generally increasing in magnitude in proportion to the number of passages administered (Royer, 2005). Marcotte and colleagues (2019) found that two- and three-passage formats, each including 16 questions per passage, produced reliable scores with coefficient alpha .78 for two- and .83 for three-passage probes. These positive reliability statistics are important because the assessment takes time to administer and the smaller the number of needed passages, the more acceptable the measure might be to those involved in planning, implementing, and evaluating RTI use.

Recent inquiry has sought to expand the potential utility of SVT by evaluating its ability to serve as a

formative measure of content comprehension. Mooney and Lastrapes (2016) evaluated SVT's effectiveness as a content measure as part of a study in which multiple assessments were administered and compared to criterion measures in science and social studies content. Measures targeting academic vocabulary (i.e., critical content monitoring; Mooney, McCarter, Russo, & Blackwood, 2013), written expression (i.e., written retell; Marcotte & Hintze, 2009), and reading comprehension (SVT) were positively correlated with state content accountability test scores. Results showed that SVT and critical content monitoring were significant predictors of achievement in fifth-grade science and social studies (Mooney & Lastrapes, 2016). One reason that SVT might be considered for inclusion in RTI frameworks is that students carry out tasks that are similar to what they are expected to do during state accountability testing—that is, students are presented with a content-area passage and asked to respond to multiple choice questions about the passage.

Although the research investigating different measures of reading and content comprehension has been examined, what is not known is whether a science content passage will be a valid indicator of course learning, similar to that originally proposed by Deno (1985), relative to CBM. The SVT is different from other CBM tools in that students are reading a full science content passage and responding to statements about that content that either confirm or diverge from the passage. Not only does this activity mimic what students will see on state achievement tests, but SVT is presented and automatically scored in an online version and can be administered across grades, thus making it easier for teachers to administer than paper-pencil tests. These features potentially increase its instructional utility (Mooney & Lastrapes, 2016).

The purpose of the present study was to determine if SVT was a significant indicator of science content comprehension, replicating the already robust criterion validity research and extending inquiry to areas of social validity, internal consistency, and test-retest reliability. The authors asserted that the present inquiry would inform the literature in two ways. First, robust correlations between a content SVT and content state accountability scores would potentially increase the viable CBM instrument options for practitioners and researchers at the upper elementary and secondary school levels. Inquiry in content area CBM is still in its infancy, particularly when compared with reading CBM (Mooney & Lastrapes, 2016). Second, increasing the instrument options might allow for greater flexibility in the design and implementation of screening and progress monitoring assessment frameworks at the upper elementary and secondary school levels. Such work could serve to more quickly identify at-risk learners for intervention. Research in upper elementary and secondary RTI frameworks is itself in its formative stages (Mooney & Lastrapes, 2018a), with Fuchs, Fuchs, and Compton (2010) suggesting that upper elementary and secondary systems be considered differently than the more traditional elementary-oriented frameworks.

The following research questions were addressed: (a) What is the distribution of SVT scores across grade levels? (b) What are demographic subgroup comparisons for SVT benchmark probe scores and criterion tests (i.e., a statewide science accountability test and a nationally-normed standardized achievement test)? (c) What is the strength of the relationship between the SVT benchmark probes and a state content accountability test? (d) What is the strength of the relationship between the SVT benchmark probes and a nationally representative achievement test? (e) What is the social validity of SVT? and (f) What are the internal consistency and test-retest reliability statistics for SVT measures with 16 (one-passage) and 32 (two-passage) test items?

Method

Participants

Participants were students enrolled in a rural district in south Louisiana. District students in the fourth (n = 147), fifth (n = 216), and sixth grades (n = 123) who completed all periodic SVT probes and had state achievement scores available to the district comprised the sample. Students' ages for the state test sample ranged from 9.4 to 13.9 years (mean = 11.6). The sample was 68% African American, 31.4% White, and less than 1% two or more races; 85% lower socioeconomic (SES) status; and 7% disabled. At the request of the participating school district, a subset of the sample (n = 76) in two of the district's schools was also administered the SAT 10th edition online abbreviated test (SAT-10). In School A (n = 19), only fifth and sixth graders were tested; in School B (n = 57) fourth, fifth, and sixth graders were tested. That sample makeup was 55% female, 94% African American, 96% low SES, and 3% disabled.

Instrumentation

Three benchmark SVT probes were administered along with two criterion and two social validity measures, one to teachers and one to students. Originally created in 1986 and revamped in 1999 (Decuir, 2012), the Louisiana Educational Assessment Program (LEAP) and the integrated LEAP (iLEAP; fifth and sixth grades) were used as the criterion measures. The science subtest of the online abbreviated SAT-10 was administered to the smaller sample only.

Predictor: SVT. The SVT probes used in the present study consisted of two passages and 16 accompanying sentences per passage. On an SVT test (see Figure 1 for an example), the participant read a passage and then was presented with four types of test sentences. Without referring to the previously read passage, the test taker selected *yes* if the meaning of the passage was preserved in the sentence read or *no* if the meaning was not preserved across 16 test sentences. The four item types were: originals (exact copy of a sentence from the passage), paraphrases (meaning preserved using similar but not the same wording), meaning changes (exact copy of the sentence with a minor change that altered the meaning), and distractors (a sentence that was not part of the reading passage and usually drawn from a

Litter is Pollution You Can See
Flesch-Kincaid Grade Level 5.8
We pollute when we add things to the environment that are harmful. Pollution can affect the air, soil, or water of an ecosystem. Pollution can make people or animals sick. Pollution can cause diseases or death. Trash on the ground is called litter. Littering is one way that people pollute. Litter harms both plants and animals. Litter can kill plants and entangle animals. Old fishing lines and nets kill many aquatic animals. Some animals might eat litter and get sick. Sea turtles are especially harmed when clear plastic sandwich bags are thrown into the ocean. The sea turtle thinks the bag is a jellyfish, which is what they eat. The bag gets stuck in the turtle's stomach. If a bag is stuck in an animal's stomach, it will eventually starve. Fishing lines, balloons, and plastic bags last a long time in the ocean. These plastic materials are not biodegradable, which means that they do not break down easily.

Item	Type	Statement
1	M	Litter is good for plants and animals.
2	D	Normally all ecosystems are naturally balanced.
3	D	Non-native species introduction can be a threat to ecosystems.
4	O	Trash on the ground is called litter.
5	P	Adding harmful things to the environment is called pollution.
6	O	Pollution can affect the air, soil, or water of an ecosystem.
7	P	Plants and animals can get sick because of pollution.
8	M	Litter could injure animals, but it does not affect plants.
9	D	If a species does not have a predator, it will grow out of control.
10	O	Old fishing lines and nets kill many aquatic animals.
11	M	Plastic materials are biodegradable, because they break down easily.
12	P	Plastic bags can get stuck in the stomach of a sea turtle.
13	P	A sea turtle will eat a plastic bag, because it looks like a jellyfish.
14	O	Some animals might eat litter and get sick.
15	D	Invasive earthworms eat different things than native earthworms.
16	M	Fishing lines, balloons, and plastic bags dissolve quickly in the ocean.

Figure 1. Example of SVT benchmark probe, sentence type, and (correct answer); M = meaning change (No), D = distracter (No), O = original (Yes), P = paraphrase (Yes).

nearby paragraph in the source material). In this study, there were four original, paraphrase, meaning-change, and distractor sentences that were randomly arranged for each passage. All SVT probes were developed by the first author with the assistance of a district science curriculum specialist. To create the probes, fourth-, fifth-, and sixth-grade texts and Louisiana grade-level expectations (Louisiana Department of Education; LDE, 2014a) were examined for relevant text content. Previous criterion validity correlations for a paper-pencil version of SVT were .46 (95% confidence intervals [95% CI] .21, .66) with iLEAP and .49 (.25, .67) with SAT-10 science in fifth grade (Mooney & Lastrapes, 2016).

Criterion: i/LEAP. The stated purpose of the LEAP and iLEAP was to measure progress toward Louisiana's academic standards in English/language arts, math, science, and social studies for all students in Grades 3 through 9 (LDE, 2014a). The science tests included multiple-choice questions. Science assessments addressed science as inquiry, physical, life, earth, space, and environmental science (LDE, 2014a, 2015). Achievement-level descriptors were unsatisfactory, approaching basic, basic, mastery, and advanced. Technical adequacy data for the i/LEAP tests were accessed from the LDE website. Reliability statistics consisted of internal consistency Cronbach's alpha correlations for the science subtests of .85 (fourth-grade test), .87 (fifth-grade test) and .88 (sixth-grade test; LDE, 2014b). State-provided validity data were described in terms of a content validity process that was not delineated (LDE, 2014b).

Criterion: SAT-10. The abbreviated form of the online SAT-10 was a standardized, norm-referenced achievement test battery that measured reading, mathematics, spelling, language, listening, science, and social studies performance for students in kindergarten through 12th grade. Pearson Education (2015) described the science test as aligned with national and state content standards. The test-derived scaled score was used in the present study. The scaled score was vertically equated across each subject test, reportedly allowing for the tracking of performance across grades (Pearson Education, 2015). The science test assessed science as inquiry, knowledge of life, physical, and earth sciences. The abbreviated battery content test consisted of 30 multiple-choice questions and was untimed. Mooney and Lastrapes (2016) reported a .64 (.44, .78) correlation with the i/LEAP science test in fifth grade. In the present study, correlations were .60 (.44, .74), .50 (.35, .71), and .61 (.37, .81) for fourth, fifth, and sixth grades, respectively.

Social validity. Teachers were presented with a social validity survey at the end of the school year. Two questions were presented: (a) In your opinion, would the information gained from the SVT procedure be helpful in informing your practice in developing course curriculum; establishing progress toward goals; measuring progress toward goals; deciding when to change instruction; communicating student progress to other school personnel, to parents, and to students and (b) How time consuming were the SVT procedures?

In May, two social validity statements were added to the SVT administered to the students. The researcher-created statements were (a) "I think the SVT can help me show what I am learning in class", and (b) "It was easy to use the SVT." Students responded using a Likert scale, with 1 = *not at all*, 2 = *a little bit*, 3 = *more than a little bit*, and 4 = *very much*. This was the third of three informal social validity surveys in the content CBM literature, none of which have been administered technical adequacy analyses.

Procedures

Students took three benchmark assessments, in October, January, and May, with a February probe given, identical to the May probe, for test-retest reliability purposes. The assessments were delivered via the Qualtrics online survey software system. The software presented the assessments to the students through a web link. The web links were placed on the district's website for the classroom teacher to guide students through the test-taking process. Students typed in their names and selected their school, grade, and teacher. Teachers then led students through standardized directions and oversaw test completion. Upon completion, the software presented correct and incorrect scores (and answers) to the students. The first SVT was piloted the spring of the school year prior to the start of the present study to determine how long it would take the students to complete. Times ranged from 15–20 minutes from directions being read by the teacher to completion of the test items.

The first author administered the initial October SVT probes with the teachers over the course of a week to facilitate fidelity of implementation. During subsequent administrations, district teachers were notified when the testing week would occur, and teachers administered the tests independently. The first author conducted

observations at each of the schools during the week of assessments, and anecdotal evidence indicated that SVT implementation went as planned. No formal fidelity checks were conducted. An assessment was administered in February and repeated in May for the final benchmark measure to calculate test-retest reliability. The large gap in time between assessments was filled with district-wide and state accountability testing and two week-long holiday breaks. Teachers administered the statewide accountability test in April of the academic year in accordance with guidelines established by the state department of education. Authors administered the SAT-10 test shortly before state testing.

Data Analysis

Following each administration, SVT data were downloaded from Qualtrics and maintained in an SPSS master file of total scores. Descriptive statistics were calculated and examined to determine the distribution of SVT benchmark scores and criterion test scores across grade levels. Using separate multiple regressions, comparisons were examined for the dichotomous demographic categories gender, education classification (special education vs. general education), and SES status (high vs. low). Race was treated as a dichotomous variable limited to White and African American students, as only 27 students were in a different racial category. Data from these 27 students were excluded from the analysis regarding race but included in the other analyses.

To examine criterion validity for the state test, students (level one) were nested within seven schools (level two), which implied a hierarchical structure to the data. To address this issue, initially a multilevel data analysis technique was used to determine the unexplained variation in the outcome variables (with state achievement test and SVT benchmark probe scores all continuous variables) across each school. The unconditional analysis of variance model (the null model) was examined without level one or level two predictors to determine if there were any unexplained variation among the schools. Unexplained variation was not found ($p > .05$ for all five tests) across the seven campuses for each of the outcome variables as well as SVT probes. Therefore, a single level analysis (correlation) was employed to analyze the data. To quantify the strength of the relationship between benchmark SVT probes and both the state accountability tests, point estimates and 95% CI of Pearson's r were calculated. To determine the social validity of the probes, descriptive statistics were calculated for the teachers' and students' responses. For student social validity, mean totals for each of the two statements were computed. For teacher social validity, the greatest percentages of responses chosen were reported.

Regarding reliability, SVT probes were analyzed for internal consistency reliability using Cronbach's alpha. Although the SVT benchmark probe was two passages with 32 corresponding test items, each individual passage (16 test items) was also examined for internal consistency to see if future probes of a single passage were feasible. Pearson's r was examined to determine the test-retest reliability of the February and May SVT probes, with both passages analyzed together and each passage analyzed separately.

Results

Demographics

Table 1 provides descriptive test data for the four SVT probes and both the state achievement and SAT-10 tests. Mean SVT probe scores ranged from 19.2 (of 32 possible correct answers) to 25.9 correct answers across three independently administered probes, with the lowest score across all grades in January. Table 2 compares the predictor and criterion test score patterns with respect to demographic and classification subgroups. Results demonstrated considerable variability. Statistical differences among subgroups for the May SVT probe scores matched those of the state achievement test in two of four cases and those of the national test in three of four cases. There was less agreement for the October and January comparisons. Across SVT probe scores, the only consistent match between all three predictor and both criterion measures was in regards to race.

Concurrent and Predictive Criterion

The results for concurrent and predictive correlations are reported in Table 4. Concurrent validity (May) ranged from .24 (95% CI .07, .40) to .55 (.43, .65) for SVT and i/LEAP/LEAP and .18 (-.27, .56) to .27 (-.21, .51) for SVT and SAT-10. Predictive SVT correlations ranged from .16 (-.02, .32) to .49 (.38, .58) with i/LEAP/LEAP and -.09 (-.45, .30) to .57 (.25, .77) with SAT-10. For the i/LEAP/LEAP correlations, all were statistically significantly correlated with the SVT probes for each month, with the exception of January for sixth grade. For SAT-10, only the October correlation with SVT in sixth grade was statistically significant.

Social Validity

Teachers. The social validity questionnaire was administered to the participating teachers at a Saturday meeting at the end of the school year, which was sparsely attended. Only 15 teachers completed the survey. Of those, the majority regarded the SVT as very helpful in improving practices in measuring progress toward goals and communicating student progress to other school personnel, parents, and students; helpful in establishing goals, measuring progress toward goals. and deciding when to change instruction; and somewhat helpful in course curriculum. Sixty-six percent thought the SVT procedure was not very or not at all time consuming, 27% said it was somewhat time consuming, and 7% (1 respondent) said it was very time consuming.

Students. To determine the social validity of SVT, means and standard deviations were calculated for the two statements, with a highest possible score of four. All grades agreed with both statements, with ease of use judged more favorably than its utility by the fourth and

Table 1

Distributions of Means, Standard Deviations, and [95% Confidence Intervals] by Grade

4th grade	*M*	*SD*	[95% CI]	Range	Skew	Kurtosis	*N*
SVT	21.5	3.9	[20.9, 22.1]	12-30	-.19	-.34	147
October							
January	19.4	4.2	[18.6, 20.1]	8-29	.02	-.10	147
May	24.8	4.9	[24.0, 25.6]	11-32	-.59	-.35	147
i/LEAP	324.9	45.3	[317.5-332.2]	100-407	-1.1	3.5	147
SAT-10	620.0	34.4	[607.0, 634.3]	551-681	.07	-.51	27
5th grade							
SVT	21.2	4.5	[20.6, 21.8]	11-31	.11	-.42	216
October							
January	19.5	4.7	[18.8, 20.2]	7-30	-.05	-.55	216
May	25.9	4.6	[25.3, 26.5]	10-32	-.83	.36	216
i/LEAP	306.6	46.1	[300.4, 312.8]	100-462	-.71	2.1	216
SAT-10	631.1	26.2	[619.0, 643.0]	584-681	.02	-.52	21
6th grade							
SVT	21.8	4.3	[21.0, 22.6]	10-32	-.29	.33	123
October							
January	19.2	4.9	[18.3, 20.1]	8-31	.30	-.37	123
May	24.1	5.6	[23.1, 25.1]	4-32	-.81	.66	123
i/LEAP	303.9	42.8	[306.3, 307.0]	100-384	-1.6	7.1	123
SAT-10	645.9	19.7	[638.2, 653.5]	612-703	.78	1.3	28

Note. M = mean, SD = standard deviation, CI = confidence interval; SVT = Sentence Verification Technique; i/LEAP = Louisiana state achievement test, SAT-10 = Stanford Achievement Test Series, Tenth Edition.

Table 2

Comparison of Criterion and Predictor with Respect to Differences in Subgroup Mean Scores (N = 486)

Assessment	*B*	SE_B	β
i/LEAP			
Male – Female	5.35	3.8	0.06
White – African American	-24.9	4.1	-0.27**
General education – Special education	-43.5	7.6	-0.23**
High SES – Low SES	-18.9	5.6	-0.14**
October SVT			
Male – Female	-0.74	0.38	-0.86*
White – African American	-1.22	0.40	-0.14**
General education – Special education	-3.09	0.75	-0.18**
High SES – Low SES	-0.77	0.56	-0.06**
January SVT			
Male – Female	-0.51	0.41	-0.06
White – African American	-2.11	0.44	-0.22**
General education – Special education	-2.39	0.81	-0.13**
High SES – Low SES	0.55	0.60	0.04
May SVT			
Male – Female	5.05	6.69	0.09
White – African American	-43.1	15.23	-0.33**
General education – Special education	-43.7	22.71	-0.24
High SES – Low SES	-5.44	18.4	-0.0
SAT – 10			
Male – Female	-0.56	0.43	-0.0
White – African American	-2.87	0.46	-0.21
General education – Special education	-4.47	0.86	-0.22
High SES – Low SES	0.11	0.63	0.0

Note. **p < .01, *p < .05. i/LEAP = Louisiana state achievement test; SVT = Sente Verification Technique. SES = socioeconomic status. First group is the reference gr analysis limited to African-American and White students ("Other," $N = 10$); B = Un coefficient, SE_B = Standard error of the coefficient; β = Standardized coefficient.

Table 3

Sentence Verification Technique Total Cronbach's Alpha Coefficient by Month and Grade

	Passage 1 (16 items)	Passage 2 (16 items)	Combined (32 items)
October			
Overall	.47	.58	.70
4th grade	.43	.49	.59
5th grade	.54	.58	.71
6th grade	.48	.64	.71
January			
Overall	.67	.69	.69
4th grade	.48	.40	.43
5th grade	.54	.51	.48
6th grade	.63	.57	.60
May			
Overall	.73	.70	.82
4th grade	.72	.69	.83
5th grade	.66	.71	.81
6th grade	.77	.64	.82

Table 4

Benchmark Correlations and [95% Confidence Intervals] of Sentence Verification Technique and Criterion by Grade

	Fourth Grade (N = 147)	Fifth Grade (N = 216)	Sixth Grade (N = 123)
i/LEAP			
SVT October	.45** [.31, .57]	.49** [.38, .58]	.48** [.33, .60]
SVT January	.27** [.11, .41]	.41** [.29, .51]	.16 [-.02, .32]
SVT May	.55** [.43, 65]	.53** [.43, .62]	.24** [.07, .40]
SAT-10			
SVT October	.28 [-.11, .58)	-.01 [-.44, .42]	.57** [.25, .77]
SVT January	-.09 [-.45, .30)	.12 [-.33, .52]	.13 [-.25, .48]
SVT May	.19 [-.21, .53)	.18 [-.27, .56]	.27 [-.21, .51]

Note. **Correlations are significant at $p < .01$, all others not significant. i/LEAP = Louisiana state achievement test, SAT-10 = Stanford Achievement Test Series, Tenth Edition.

fifth graders and the opposite for the sixth graders. For statement 1, "I think the SVT can help me show what I am learning in class," results were M = 3.04, SD = .94 for fourth grade (n = 194); M = 3.06, SD = .96 for fifth grade (n = 267); and M = 3.15, SD = .99 for sixth grade (n = 166). For the statement, "It was easy to use SVT," results were M = 3.12, SD = 1.0 for fourth grade; M = 3.31, SD = .89 for fifth grade; and M = 3.12, SD = 1.0 for sixth grade.

Internal Consistency and Test-Retest Reliability

For internal consistency, Cronbach's alpha was calculated for both passages separately and together (see Table 3). Alpha coefficients for the combined passages for the entire sample were .70 for October, .69 for January, and .82 for May and ranged from .43 to .83 overall by grade. Single-probe statistics ranged from .40 to .77 with May showing the strongest reliability values by individual passage overall (range .66–.77). The test-retest reliability of SVT probes for February and May were analyzed using Pearson's r statistic and 95% CIs overall and by grade. Passages were analyzed individually as well as collectively. The overall test-retest correlation, calculated for all grades with no missing data (N = 620) between February and May SVT probes, was .63 (95% CI .58, .68). For the sample by grade, the results for fourth grade (n = 193) were .70 (.62, .77); the results for fifth grade (n = 265) were .68 (.61, .74); and the results for sixth grade (n = 162) were .68 (.59, .76).

Discussion

At-risk learners can benefit from effectively designed and delivered RTI systems. Research into RTI system elements for older at-risk students is ongoing, with questions targeting content-oriented frameworks particularly prescient. The present study evaluated the utility of an online and abbreviated form of an assessment (SVT) that previously was shown to be an effective measure of reading comprehension (Marcotte & Hintze, 2009; Royer, 2005). Validity and reliability questions were addressed with a large, multigrade sample of public school students. Results on distribution patterns, reliability, and validity are discussed. Study limitations and implications are addressed.

In evaluating the tenability of a new purpose for an existing assessment instrument, it makes sense to look for similarities in the distribution of scores among predictor and criterion or comparison instruments, with the goal that an assessment would be sensitive enough to accurately detect those students who are at risk for academic difficulties. In the present data, such patterns were not readily discernible. As expected and desired from a progress monitoring determination perspective, the mean scores for the May benchmark were higher than those of October (see Table 1). That pattern was evident across all grades, which itself is informative in a content area literature that has largely targeted single-grade-level-student participation. However, the mean scores did not show a clear pattern of higher scores as the grade level of a student increases, an array that was evident with scores of the standardized SAT-10 measure. Moreover, there was variability in the performance scores, with lower mean scores in January from October across all grades. The lower scores may have been due to lower overall reliability statistics for the January probe than those of October or May.

There was also variability in the comparison of subgroup scores from criterion to predictor measure. In the content literature, there has been some measure of agreement in the few times that comparisons have been made. In the structured formative assessment (e.g., CBM) literature, measures of academic language such as vocabulary matching (Espin & Deno, 1994-1995) and critical content monitoring (Mooney et al., 2013) have demonstrated the ability to match criterion measure patterns. For example, Mooney, McCarter, Schraven, and Haydel (2010) reported that there were comparable statistical patterns in 9 of 10 cases for a sixth-grade social studies sample using vocabulary matching and state accountability test scores. In a Grades 4–5 sample, Mooney and Lastrapes (2018a) reported comparable statistical patterns with state accountability test scores for critical content monitoring relative to gender, disability status, and socioeconomic status but not for race. For SVT, there was consistency in the comparison of differences in scores across race. In all six comparisons, SVT statistical patterns mirrored those of criterion measures. One inconsistent pattern example related to disability status. That is, SVT was consistent with predictor variables for disability status in October and January but not May. These data call into question the viability of the SVT to be able to detect students at risk for learning disabilities, particularly when compared to CBM measures that have better track records of pattern matching.

Criterion-related validity correlation magnitudes were stronger for state-level comparisons and, overall, generally low to moderate in strength across grades (see Table 4). Overall, linear relations between measures ranged from .16 (95% CI -.02, .32) to .55 (.43, .65) in nine comparisons, with six of those in the moderately strong range. State-level concurrent validity correlations ranged from .24 (.07, .40) in sixth grade to .55 (.43, .65) in fourth grade. Predictive magnitudes ranged from .16 (-.02, .32) to .49 (.38, .58), with descriptively stronger correlations for the fall benchmarking period. Overall correlations between the SVT and SAT-10 national test were considerably weaker in magnitude, with only one of nine linear relationships (i.e., fall benchmark, sixth grade, .57 [.25, .77]) reported in the moderately strong range.

Concurrent validity correlations were generally lower in magnitude than those reported in the structured formative assessment literature. In science content, linear relationships for vocabulary matching (Espin et al., 2013) and content maze (Johnson et al., 2013) have been in the .6 to .7 range, while critical content monitoring has evidenced .45 to .55 findings in state-level comparisons (Mooney et al., 2013; Mooney & Lastrapes, 2018a) and .67 with a national test (Mooney & Lastrapes, 2018a). The few predictive benchmark magnitudes have generally

been lower than the concurrent associations, with correlations of .46 reported for content maze (Johnson et al., 2013) and .38–.58 for critical content monitoring (Mooney & Lastrapes, 2016, 2018a).

Social validity findings added to the small literature base in content area CBM. In the present study teachers were surveyed, a first in the literature, with a majority of those completing surveys indicating that the assessment has instructional utility. Students perceived the tool as potentially helpful to learning and easy to use. Student perspectives were similar to those surveyed in Mooney et al. (2013) targeting the critical content monitoring probe.

The SVT probe displayed moderately strong internal consistency correlations across October and May administrations of two-probe passages overall, but were more variable when analyzed by grade. Cronbach's alpha coefficients for two-passage probes ranged from .43 to .83 across grades, with only 3 of 9 coefficients above .80. Findings were comparable to previous findings for SVT reading and listening comprehension probes that ranged in length from two to six passages (e.g., Marcotte et al., 2019; Royer, Sinatra, & Schumer, 1990). In the literature, SVT alpha coefficients have varied from .5 to .9 across studies and generally increased in magnitude as the length of the probe expanded (Royer, 2005). Internal consistency results were less favorable than those reported for an adaptation of SVT known as statement verification for science that was evaluated using students in Grades 7 and 8 (Ford & Hosp, 2017). Test-retest correlations for the two-passage probe (administered in February and May) were moderately strong in magnitude across grades.

Limitations

Several limitations likely impacted the evaluation of study findings. First, although the large, multigrade sample may be representative of more urban populations, the sample was not representative of the larger United Status population demographically, thereby limiting the potential generalizability of findings. Moreover, the academic achievement of the smaller sample of students who were administered the SAT-10 science test was depressed, possibly reducing the magnitudes of the correlations with SVT. Second, because the SVT passages were adapted from classroom texts, which have been demonstrated to use higher level vocabulary, the readability levels of the SVT passages were all challenging for sample grade levels. That differed from previous probe development processes in which multiple-passage probes had readability levels below, at, and above the participant levels. The grade readability range for the Grades 4–6 population of students may have depressed scores as well as limited the generalizability of findings to the SVT literature. There were no technical adequacy checks for the social validity questionnaires, which may impact results for the teachers' and students' opinions of the measures. Finally, the particular probe administration program the authors utilized (Qualtrics) did not allow for immediate teacher access to useable data, which may have impacted subsequent SVT scores over the course of the study.

Implications for Research and Practice

With study findings and limitations described, we focus implications discussion on SVT's potential inclusion in school-based RTI frameworks and research that may be relevant to informing that circumstance. Effective implementation of RTI systems has the potential to improve academic performance for at-risk learners.

As it stands now, there is a literature base that supports SVT's inclusion in upper elementary and secondary school literacy assessment frameworks. The technical adequacy literature summarized by Royer (2005) and extended by Marcotte and Hintze (2009) is evidence of that. With supportive validity and reliability data, SVT can be considered an instrument to document school-wide student performance as part of the first tier of RTI frameworks in reading. In RTI systems, all students' performance is evaluated in a standardized fashion periodically, with those deemed at risk or delayed receiving supplementary intervention support and more frequent testing to evaluate the success of the extra school effort. The SVT assessment has a history of documenting reading achievement through paper-pencil administration procedures. As a result of the present study, there is also reason to believe that an online delivery system might facilitate extending SVT into upper elementary and secondary school RTI frameworks. Teachers seemed generally receptive to SVT, indicating that it was not overly time consuming and could assist in instructional decision-making. Teachers, with very little training, also managed to direct the assessment implementation process.

Still, our hypotheses asserted that SVT probes could be extended into science courses and provide robust technical adequacy findings as have been demonstrated in reading. We do not believe that those hypotheses were supported. The concurrent and predictive correlations with meaningful criterion were generally inferior to those of other instruments, such as vocabulary matching, critical content monitoring, and content maze, a finding that is not new (Johnson et al., 2013; Mooney & Lastrapes, 2016, 2018b). There was also not the consistency of patterns in scoring evident for the content SVT probe as there has been for vocabulary matching and critical content monitoring (Mooney & Lastrapes, 2018a; Mooney, McCarter, Schraven, & Callicoatte, 2013). As the evaluation of robust interventions to address the reading concerns of at-risk students continues, so too should the investigation of meaningful formative assessment (e.g., CBM) tools for upper elementary and secondary school students, particularly in the area of content area courses.

References

Barth, A. E., Tolar, T. D., Fletcher, J. M., & Francis, D. (2014). The effects of student and text characteristics on the oral reading fluency of middle-grade students. *Journal of Educational Psychology*, *106*, 162–180. doi:10.1037/a0033826

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

Boudreaux-Johnson, M., Mooney, P., & Lastrapes, R. L. (2018). An evaluation of close reading with at-risk fourth-grade students in science content. *Journal of At-Risk Issues*, *20*(1), 27–35. https://eric.ed.gov/?id=EJ1148245

Conoyer, S. J., Ford, J. W., Smith, R. A., Mason, E. N., Lembke, E. S., & Hosp, J. L. (2018). Examining curriculum-based measurement screening tools in middle school science: A scaled replication study. *Journal of Psychoeducational Assessment,* Advance online publication. doi:10.1177/0734282918803493

Decuir, E. L. (2012). *Louisiana Educational Assessment Program* (LEAP): *A historical analysis of Louisiana's high stakes testing policy*. (Doctoral dissertation, Georgia State University). Retrieved from https://scholarworks.gsu.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1101&context=msit_diss

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*, 219–232. doi:10.1177/001440298505200303

Espin, C. A., Busch, T., Lembke, E. S., Hampton, 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

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. https://eric.ed.gov/?id=EJ513524

Ford, J. W., Conoyer, S. J., Lembke, E. S., Smith, R. A., & Hosp, J. L. (2018). A comparison of two content area curriculum-based measurement tools. *Assessment for Effective Intervention*, *43*(2), 121–127. doi:10.1177/1534508417736753

Ford, J. W., & Hosp, J. L. (2017). Statement verification for science: Theory and examining technical adequacy of alternate forms. *Exceptionality.* Advance online publication. doi:10.1080/09362835.2017.1375410

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, *57*, 488-500.

Fuchs, D., & Fuchs, L. S. (2019). On the importance of moderator analysis in intervention research: An introduction to the special issue. *Exceptional Children*, *85*, 126–128. doi:10.1177/0014402918811924

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2010). Rethinking response to intervention at middle and high school. *School Psychology Review*, *39*(1), 22–28.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). New York, NY: Guilford.

Johnson, E. S., Semmelroth, C., Allison, J., & Fritsch, T. (2013) The technical properties of science content maze passages for middle school students. *Assessment for Effective Instruction*, *38*, 214–223. doi:10.1177/1534508413489337

Louisiana Department of Education. (2015). *i/LEAP interpretive guide, grades 3–8, science and social studies* (Spring 2015). Baton Rouge: Author.

Louisiana Department of Education. (2014a). *LEAP interpretive guide, grades 4–8, science and social studies*. Baton Rouge: Author.

Louisiana Department of Education. (2014b). *LEAP 2014 operational technical summary.* Baton Rouge: Author.

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology*, *47*(5), 315–335. doi:10.1016/j.jsp.2009.04.003

Marcotte, A. M., Rick, F., & Wells, C. S. (2019). Investigating the reliability of the Sentence Verification Technique. *International Journal of Testing*, *19*(1), 74–95. doi:10.1080/15305058.2018.1497636

Mooney, P., & Lastrapes, R. E. (2016). The benchmarking capacity of a general outcome measure of academic language in science and social studies. *Assessment for Effective Intervention*, *41*, 209–219. doi:10.1177/1534508415624648

Mooney, P., & Lastrapes, R. E. (2018a). Conceptual replications of the critical content monitoring general outcome measure in science content. *Assessment for Effective Intervention*. Advance online publication. doi:10.1177/1534508418791733

Mooney, P., & Lastrapes, R. E. (2018b). Replicating criterion validity in science content for the combination of critical content monitoring and sentence verification technique. *Assessment for Effective Intervention*. Advance online publication. doi:10.1177/1534508418758362

Mooney, P., Lastrapes, R. E., Marcotte, A. M., & Matthews, A. (2016). Validity of two general outcome measures of science and social studies achievement. *Specialusis Ugdymas/Special Education*, *34*(1), 145–188. http://socialwelfare.eu/index.php/SE/article/view/253

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*(4), 249–260. doi:10.1177/1534508413488794

Mooney, P., McCarter, K. S., Schraven, J., & Haydel, B. (2010). The relationship between content area general outcome measurement and statewide testing in world history. *Assessment for Effective Intervention*, *35*, 148–158. doi:10.1177/1534508409346052

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*, 85–100. doi:10.1177/001440291308000104

Nation's Report Card. (2017). U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://www.nationsreportcard.gov/reading_math_2017_highlights/

Pearson Education (2015). *Stanford Achievement Test series, online abbreviated form* (10th ed.). https://images.pearsonassessments.com/Images/PDF/Webinar/Stanford_Testing_Info_Packet1272011.pdf

Royer, J. M. (2005). Uses for the sentence verification technique for measuring language comprehension. *Progress in Education*. Retrieved from https://www.readingsuccesslab.com/wp-content/uploads/2015/07/Svt-Review.pdf

Royer, J. M., Hastings, C. N., & Hook, C. (1979). A sentence verification technique for measuring reading comprehension. *Journal of Literacy Research*, *11*(4), 355–363. doi:10.1080/10862967909547341

Royer, J. M., Sinatra, G. M, & Schumer, H. (1990). Patterns of individual differences in the development of listening and reading comprehension. *Contemporary Educational Psychology*, *15*, 183–l96. doi:10.1016/0361-476X(90)90016-T

Vaughn, S., Roberts, G. J., Miciak, J., Taylor, P., & Fletcher, J. M. (2019). Efficacy of a word- and text-based intervention for students with significant reading difficulties. *Journal of Learning Disabilities*, *52*(1), 31–44. doi: 10.1177/0022219418775113

Wexler, J., Reed, D. K., Barton, E. E., Mitchell, M., & Clancy, E. (2018). The effects of a peer-mediated reading intervention on juvenile offenders' main idea statements about informational text. *Behavioral Disorders*, *43*, 290–301. doi:10.1177/0198742917703359

Authors

Renée E. Lastrapes, PhD, is an Assistant Professor in the Department of Educational Research & Assessment at the University of Houston–Clear Lake. Dr. Lastrapes' research interests are assessment for students with specific learning disabilities and evidence-based behavioral and academic interventions for teachers working with students with emotional and behavioral disorders.

Paul Mooney, PhD, is a Professor and Director of Special Education Programs in the School of Education at Louisiana State University. His research interests include the utility and makeup of practitioner journal articles and structured formative assessment and academic interventions for students with or at risk for academic or behavioral disabilities.