TEST-RETEST RELIABILITY OF VOCABULARY MATCHING IN SIXTH-GRADE WORLD HISTORY

Paul Mooney[1], Jodie Schraven
Louisiana State University, USA

Ben Cox
West Baton Rouge Parish School Board, USA

Abstract. ***Background***. The present study was designed to extend reliability research on a content-area curriculum-based measurement tool known as vocabulary matching. ***Purpose***. Test-retest reliability of vocabulary matching was evaluated with a diverse sample of 39 sixth-grade students from a rural middle school in the southeastern United States. ***Material and Methods***. Students were administered the same five-minute probe on two separate occasions with five instructional days between administrations. ***Results***. A correlation of $r = .91$ (95 % confidence intervals = .83, .95) was determined, providing evidence of stability for the content-area progress-monitoring tool. ***Conclusions***. Findings add to a growing body of technical adequacy research supporting vocabulary-matching's utility in measuring performance and progress in content-area courses. Research implications and future directions are discussed.

Keywords: Vocabulary Matching, Progress Monitoring, Secondary Schools, Content Areas

TEST-RETEST RELIABILITY OF VOCABULARY MATCHING IN SIXTH-GRADE WORLD HISTORY

If K-12 response-to-intervention system frameworks are going to function in public school settings, then it is critical that educators continue to develop and validate progress monitoring assessment systems. School-wide progress monitoring systems manage elementary school

[1] Address for correspondence: 219A Peabody Hall,Baton Rouge, LA 70803; (225) 578-2360, (225) 578-9135; e-mail: pmooney@lsu.edu.

student proficiency in reading, writing, and mathematics skills and secondary school student proficiency of content-area courses such as those in the social and natural sciences. While validation research in elementary school assessment tools is extensive [see, for example, Marston (1989) and Wayman, Wallace, Wiley, Ticha, & Espin (2007) for summaries of curriculum-based measurement studies in reading], the same cannot be said for secondary school instruments.

The Research Institute on Progress Monitoring (n.d.) has suggested that curriculum-based measures in the area of reading, writing, and content-area learning could serve as potentially useful components of a literacy assessment framework at the secondary school level. Specifically, the Institute suggests that the traditional maze curriculum-based measurement tool be used to assess reading progress. For writing progress, the Institute suggests use of a five-minute writing sample that follows a narrative prompt. For measuring progress in content-area courses, the Institute suggests use of a five-minute matching task that targets vocabulary relevant to the respective content course. A small body of technical adequacy research summarized by Busch and Espin (2003) has supported the reliability and validity of the content-area instrument known as vocabulary matching. The purpose of the present study was to extend the reliability research on this instrument.

VOCABULARY MATCHING

Vocabulary matching (Espin & Deno, 1993; Espin & Foegen, 1996) is a curriculum-based measurement (CBM; Deno, 1985) instrument designed to assess both performance and progress in content-area courses. It is one of two approaches that have been described in the literature for measuring content-area learning (Espin & Tindal, 1998), with the other known as concept maze (Ketterlin-Geller, McCoy, Twyman, & Tindal, 2006; Twyman & Tindal, 2007). Vocabulary matching's present format incorporates equivalent paper-pencil probes that students are administered for five minutes (Mooney, Benner, Nelson, Lane, & Beckers, 2007). Probe content includes randomly selected vocabulary terms and definitions that are taken from teacher notes and textbook materials, including glossaries (Espin, Busch, Shin, & Kruschwitz, 2001; Mooney, McCarter, Schraven, & Haydel, in press).

Probes consist of 20 terms and associated definitions plus two distracter definitions. A student is expected to complete as many matches between terms and definitions as he or she can within the time frame. Scoring consists of recording the number of correct matches.

Technical adequacy research on vocabulary matching has generally supported its utility in applied settings. For example, at least two criterion-related validity studies have been published in the peer-reviewed literature. One, by Espin et al. (2001), compared scores on vocabulary-matching probe scores for seventh-grade students with student' grades and social studies subtest score on the *Iowa Test of Basic Skills* (ITBS; Hoover, Hieronimus, Frisbie, & Dunbar, 1992). While correlations with student grades for the 58 participants were generally low to moderate (i.e., r = .27 to .51), the degree of association between probe scores and the ITBS subtest were moderately strong (r = .56 to .64). A more recent study by Mooney et al. (in press) found a strong correlation *(r* = .70) between a vocabulary-matching probe in sixth-grade world history and the Louisiana statewide social studies accountability test for 146 participants from a diverse middle school setting. Moreover, vocabulary-matching achievement patterns related to gender, race, ethnicity, socioeconomic status, and exceptionality status were nearly identical to those indicated on the statewide social studies test.

In terms of reliability, researchers have examined alternate-form reliability in two studies. Espin et al. (2001) compared the relationships of 11 equivalent weekly probes by comparing adjacent measures. For example, Probes 1 and 2 scores were correlated, as were scores of Probes 2 and 3, Probes 3 and 4, and so forth. In all, 10 adjacent probe correlations were calculated along with a mean score correlation. Adjacent probe correlations ranged from r = .58 to .87, with r = .70 the mean score. When adjacent probe scores were combined (e.g., Probe 1 with 2) and then compared with adjacent combinations (e.g., 1 and 2 with 3 and 4), alternate-form reliabilities increased to a mean of r = .78 and a range of r = .70 to .85. Single-measure reliabilities in Mooney et al. (in press) ranged from r = .76 to .82 across three equivalent measures. Additionally, interscorer reliability was strong for the vocabulary-matching probes, with correlations ranging from r = .994 to .998. Scoring agreement for participant total scores across probes occurred 88.1 percent of the time (range 85.2-91.7 %).

Study Rationale

If, as the Research Institute on Progress Monitoring suggests, vocabulary-matching curriculum-based measures can be included as part of middle and high school academic progress monitoring systems in content-area courses, then considerably more technical adequacy research is warranted. Reliability of the instrumentation is one such area where research can be targeted. In addition to alternate-form and interscorer reliability, test-retest reliability stands as an index of stability over time (Salvia, Ysseldyke, & Bolt, 2007). There has been very little test-retest reliability research conducted relative to vocabulary matching. In discussing their findings, Espin et al. (2001) indicated positive test-retest research findings in Espin and Foegen (1996). However, no data were reported or discussed in Espin and Foegen that related to test-retest reliability; rather, the focus of that study was on the predictive ability of a researcher-created vocabulary-matching measure. Therefore, the present study aimed to report specifically on the stability of an instrument designed to measure performance and progress in sixth-grade world history content. It was hypothesized that strong correlations would be determined.

METHOD

Participants and Setting

Sixth-grade students (N = 43) enrolled in a rural public middle school (Grades 6-8) located near a metropolitan city in south Louisiana participated in the study. Student population characteristics in the world history course were as follows: (a) 56 percent female; (b) 51 percent Caucasian, 47 percent African American, and 2 percent Hispanic; and (c) 79 percent free or reduced lunch status. Fourteen percent of the participants (n = 6) were identified with disabilities, including learning disabilities (n = 4), speech-language impairment (n = 1), and emotional disturbance (n = 1). None of the participants were identified as gifted or talented.

Participants were all taught by one social studies teacher, the third author, who oversaw administration of both probes included in the present study. The male teacher was in his sixth year of teaching at the time the assessments were administered and had earned a bachelor's degree

and secondary education certification with a concentration in social studies. He had previously received a rationale for formative assessment in social studies classes and been trained in the assessment's administration by the first author. Moreover, the teacher assisted in the probe's content development process.

Measure

Two identical researcher-developed vocabulary-matching probes for sixth-grade world history were administered to participants. The probe development process evolved from procedures described in previous studies (e.g., Espin et al., 2001; Mooney et al., 2008). It was adapted in an effort to make the technology more practical and relevant to multiple school district settings. First, in order to develop as broad a textbook content base as possible, the first author requested examiner copies of all of the sixth-grade social studies textbooks (n = 5; Boehm, Hoone, McGowan, Miramontes, & Porter, 2002; Boyd et al., 2008; Burstein & Shek, 2006; Prentice Hall, 2008a, b) that were approved for use in Louisiana public schools by the state department of education. Second, an Excel file was created that included all glossary terms from each text along with accompanying definitions. The Excel file was created by choosing one textbook, recording all glossary terms and definitions, and then moving to the next textbook and adding in all additional terms and definitions that were included in the second text and not in the first. That process was replicated until all five state-approved textbooks had been reviewed. All glossary types (i.e., subject, place, people) were reviewed for potential terms and definitions. Definitions were copied verbatim unless they were more than 15 words in length or included the vocabulary term in the definition. In the case of definitions in excess of 15 words, the definitions were shortened to 15 words or less by removing words deemed unnecessary or reworking sentences. For cases in which the definition included the term, the definition was revised to remove the listed term and insert a synonym in its place. The completed Excel database included 977 terms and accompanying definitions.

Second, a university history teacher educator with previous secondary school teaching experience reviewed the list of terms in an effort to

align the list with the Louisiana Comprehensive Curriculum (Louisiana Department of Education, 2008). The sixth grade curriculum spans the years 4000 B.C. to 1600 A.D. Terms in the textbook that were deemed not relevant to that time period by the content expert were removed. Additionally, terms/definitions included in the state's curriculum and an accompanying assessment guide (Louisiana Department of Education, 1998) but missing from the Excel file were added to the list. After the content expert review, the Excel file contained 901 terms. Third, the content expert and eight practicing middle school social studies teachers rated all 901 terms from 1 to 3, with 1 considered most important for students to know and 3 considered least important content for sixth grade students to know. Each term's scores were summed and an average rating was calculated for each term based upon the nine teachers' ratings. All terms below an average score of 1.5 on the 3-point scale were deemed priority terms for sixth grade world history and included as potential content for the probes. The new database, then, consisted of 245 terms and definitions. Finally, SAS software code was developed to create 40 randomly generated probes. Each probe was similar to the form of probes described in the literature. Specifically, each probe included 20 terms listed alphabetically and 22 definitions arranged randomly. Technical adequacy data for vocabulary matching were detailed in the introduction (Espin et al., 2001; Mooney et al., 2008).

Procedures

Both probes were administered by the students' teacher midway through the fall semester of the 2008-09 school year. The two probes were administered on separate days, with five instructional days between administrations. Vocabulary-matching protocols were scored separately by the first and second authors. Two scores for each probe were entered into a database and checked to ensure accurate data entry. Data were entered into an Excel spreadsheet by the first author, who re-scored any probe in which there were scoring differences between the two scorers. Finally, the correct score for each probe was then entered into a third column in the database by the second author and checked by the first author prior to data analyses. Data in these third columns were used for all analyses. Interscorer agreement involved the correlation of two

separate scores per probe. Descriptive analyses included the calculation of mean scores and standard deviations for each probe. Relational analyses included the calculation of a Pearson correlation between the two mean scores. A 95 percent confidence interval using a Fisher's z' transformation was also reported for the correlation.

RESULTS

Interscorer agreement correlations for Probes A and B were $r = .996$ and .995, respectively. A reliability coefficient for scorers was used in estimating the extent to which there can be generalization to different scorers because it related directly to other reliability indicators (Salvia et al., 2007).

Table 1 provides descriptive and relational test data. Mean scores for both assessments of the identical probe were very comparable, as were the range of scores. The Pearson correlation coefficient, which measured the strength of the linear relationship between the two scores was statistically significant ($p < 0.01$) and very strong in magnitude, even when 95 percent confidence intervals were taken into consideration.

Table 1. *Test-Retest Reliability Statistics for World History Scores*

Measure	N	Mean	SD	Low	High	r	95 % CI
Vocabulary Matching A	39	9.74	4.1	0	18	.91*	[.83, .95]
Vocabulary Matching B	40	9.63	3.4	1	20		

Note: 95 % CI = 95 percent confidence interval for the mean score. Vocabulary-matching scores constitute the number of correct matches.

* $p < .01$.

DISCUSSION

The purpose of the present study was to extend the technical adequacy research to test-retest reliability for a curriculum-based measure of content-area learning known as vocabulary matching. Findings lend support to the stability of the instrument as a measure of

performance. First, the correlation between two administrations of the same probe, which occurred about two months into the school year, was above $r = .90$. The $r = .91$ degree of association reported is important because Salvia et al. (2007) have suggested that reliabilities at or above $r = .90$ are "demanded" (p. 141) if individual decision-making (e.g., additional interventions or services) is warranted. Intervention-oriented decisions would certainly be likely if the vocabulary-matching measure was incorporated into a secondary school response-to-intervention framework. Second, a closer look at the probes' mean scores reveals expected patterns. Specifically, not only were the overall mean scores comparable (see Table 1), but scores for those with and without disabilities were similar as well. Mean scores for those with disabilities, albeit small in number (i.e., 4 participants for Probe A and 6 for Probe B), were 2.50 and 2.67, respectively, whereas the mean scores for those without disabilities were 10.57 and 10.85, respectively. Overall, findings appear to strengthen claims that vocabulary matching is a reliable CBM in social studies content. Data demonstrate strong reliabilities across time (i.e., test-retest) and scorers (i.e., interscorer) and moderate reliability across items (i.e., alternate-form). This is in addition to favorable criterion-related validities reported with standardized instruments (Espin et al., 2001; Mooney et al., in press), and evidence demonstrating that the instrument is sensitive to differences in individual student growth over time (Espin, Shin, & Busch, 2005). Findings also align well with test-retest reliability data published relative to curriculum-based measures of reading (Marston, 1989; Wayman et al., 2007), which have often met the $r \geq .90$ threshold of Salvia et al.

IMPLICATIONS

The major implication emanating from the present study in the context of potential response-to-intervention frameworks in content-area courses is that continued study of assessment systems at the secondary school level appears warranted. One line of research relates to technical adequacy studies, what Fuchs (2004) has termed Stage 1 studies of an instrument's static score technical features. The present body of research provides preliminary support for the reliability and validity of vocabulary

matching, but only in middle school social studies content. Application of technical adequacy research across the social and natural sciences is warranted at the secondary school level if comprehensive assessment systems are going to be implemented in middle and high school settings. A second line of research relates to vocabulary-matching's ability to measure academic growth, what Fuchs has termed Stage 2 (i.e., slope technical feature) research. Espin et al. (2005) demonstrated that the instrument is sensitive to inter-individual differences. Considerably more research is needed across content areas.

A third line of research in the area of secondary response-to-intervention frameworks might determine whether or not the vocabulary-matching measure has sufficient sensitivity and specificity to serve as a diagnostic screening tool similar to other CBM assessments. For example, an extensive body of research (e.g., Mooney et al., 2008; Silberglitt & Hintze, 2005) has indicated that oral reading fluency can accurately predict passage or failure rates on statewide accountability instruments. Similar questions could be addressed relative to vocabulary matching across a number of content-area courses in the social and natural sciences. A fourth line of research might involve comparing the effectiveness for content-area courses for vocabulary matching and concept mazes (e.g., Ketterlin-Geller, McCoy, Twyman, & Tindal, 2006). Twyman and Tindal (2007), for example, reported moderate test-retest correlations (r = .48 to .58) for two sixth grade attribute-oriented maze probes. One or both of these instruments may prove to be useful to educators applying curriculum-based measurement technologies to secondary school subject matter. Finally, a body of research that connects assessment and intervention systems is warranted. Educators need research-based practices to apply in content-area classrooms if students identified as at risk for academic failure are to be provided the tools to be able to succeed in particular subjects. Limited vocabularies are often characteristic of at-risk learners, including students with disabilities, students whose parents have limited financial means, and English language learners (Baker, Kame'enui, Simmons, & Simonsen, 2007). Such studies would likely fall under Fuchs' (2004) Stage 3 classification of research concerning instructional utility.

LIMITATIONS

The following three limitations also warrant the continuation of validation and application research. First, the sample size, though diverse, was small. Similar findings with larger sample sizes would provide stronger evidence of the measure's stability. Second, as has been the case in all of the previous published vocabulary-matching validation research, the studies have been conducted by one teacher. Comparable findings incorporating multiple teachers in multiple settings would again add scientific weight to researcher claims of the instrument's stability. Finally, the present study's findings involved administration of one probe. Future studies could involve test-retest administrations of multiple probes.

CONCLUSION

The present findings provide continued support for vocabulary matching as a reliable and valid measure of performance in social studies and other content courses. A strong correlation between two administrations of the same probe provided evidence for the stability of the instrument when utilized in an applied setting. Researchers and practitioners at the secondary school level have reason to consider vocabulary matching as a potentially viable piece of the response-to-intervention framework puzzle.

References

Baker, S., Kame'enui, E. J., Simmons, D., & Simonsen (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*. Upper Saddle River, NJ: Pearson.

Boehm, R. G., Hoone, C., McGowan, T. M., Miramontes, O. B., & Porter, P. H. (2002). *The world*. Orlando: Harcourt Brace & Company.

Boyd, C. D., Gay, G., Geiger, R., Kracht, J. B., Ooka Pang, V., Risinger, C. F., et al. (2008). *The world: Louisiana edition*. Boston: Pearson.

Burstein, S. M., & Shek, R. (2006). *World history*. Orlando: Holt, Rinehart, and Winston.

Busch, T. W., & Espin, C. A. (2003). Using curriculum-based measurement to prevent failure and assess learning in the content areas. *Assessment for Effective Intervention*, *28*(3&4), 49-58.

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

Espin, C. A., Busch, T. W., Shin, J., & Kruschwitz, R. (2001). Curriculum-based measurement in the content areas: Validity of vocabulary-matching as an indicator of performance in social studies. *Learning Disabilities Research & Practice*, *16*, 142-151.

Espin, C. A., & Deno, S. L. (1993). Performance in reading from content-area text as an Indicator of achievement. *Remedial and Special Education*, *14*(6), 47-59.

Espin, C. A., & Foegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children*, *62*, 497-514.

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.

Espin, C. A., & Tindal, G. (1998). Curriculum-based measurement for secondary students. In M. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 214-253). New York: Guilford.

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

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1992). *Iowa test of basic skills*. Chicago, IL: Riverside Publishing Company.

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.

Louisiana Department of Education (1998). *Louisiana Educational Assessment Program (LEAP) for the 21st Century: Teachers' guide to statewide assessment in social studies*. Baton Rouge, LA: Author.

Louisiana Department of Education (2005*). iLEAP assessment guide: English language arts, math, science, and social studies: Grade 6.* Baton Rouge, LA: Author.

Louisiana Department of Education (2008*)*. Comprehensive curriculum: Grade six social studies. Baton Rouge, LA: Author. Downloaded from: http://www.doe.state.la.us/lde/saa/2108.html.

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

Mooney, P., Benner, G. J., Nelson, J. R., Lane, K. L., & Beckers, G. (2007). Standard protocol and individualized remedial reading interventions for secondary students with emotional and behavioral disorders. *Beyond Behavior*, 3-9.

Mooney, P., McCarter, K. S., Schraven, J., & Haydel, B. (in press). The relationship between content area GOM and statewide testing in world history, *Assessment for Effective Intervention*.

Mooney, P., McCarter, K. S., Schraven, J., Hintze, J. M., Mooney, E., Landry, D., et al. (2008). Further evidence of oral reading fluency's utility in predicting statewide student reading proficiency. *International Journal of Psychology: A Biopsychosocial Approach.*

Prentice Hall (2008a). *History of our world*. Boston: Pearson.

Prentice Hall (2008b). *History of our world: The early years*. Boston: Pearson.

Research Institute on Progress Monitoring (n.d.). *Resources: Curriculum-based measurement at the secondary level*. Downloaded Dec. 2, 2008, from http://progressmonitoring.org/RIPMProducts2.html#cbm_secondary

Salvia, J., Ysselydyke, J., & Bolt (2007). *Assessment in special and inclusive education* (10th ed.). Boston: Houghton Mifflin.

Silberglitt, B. & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23,* 304-325.

Twyman, T., & Tindal, G. (2007). Extending curriculum-based measurement into middle/secondary schools: The technical adequacy of the concept maze. *Journal of Applied School Psychology*, *24*(1), 49-67.

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

ŽODYNO ATITIKMENŲ UŽDUOTIES TESTO – RETESTO PATIKIMUMAS VERTINANT PASAULIO ISTORIJOS ŽINIŲ PROGRESĄ ŠEŠTOJE KLASĖJE

Paul Mooney, Jodie Schraven, Ben Cox

Santrauka. ***Tyrimo tikslas***. Šio tyrimo tikslas – plėtoti ugdymo turinio įsisavinimo kokybės vertinimo instrumento, žinomo kaip žodyno atitikmenų užduotis, patikimumo tyrimus. ***Metodika***. Žodyno atitikmenų užduoties testo – retesto patikimumas buvo įvertintas ištyrus 39 šeštos klasės mokinius, besimokančius kaimo vietovės vidurinėje mokykloje, JAV pietryčiuose. Mokiniai du kartus atliko tą pačią penkių minučių užduotį. Tarp užduočių buvo penkių dienų kontrolinis intervalas. ***Rezultatai***. Rezultatai atskleidė, kad turinio kokybės įsisavinimo rodiklių įverčiai yra stabilūs, gauta koreliacija tarp matavimų buvo stipri r =0,91 (95 % pasikliautinis intervalas nuo 0,83 iki 0,95). ***Išvada***. Tyrimo rezultatai papildo žodyno atitikmenų užduoties techninio adekvatumo mokslinius tyrimus bei sustiprina įrodymus, kad šis ugdymo turinio įsisavinimo vertinimo metodas yra vertingas. Straipsnyje plačiau aptariama tyrimo rezultatų reikšmė ir tolesnės tyrimų kryptys.

Pagrindiniai žodžiai: žodyno atitikmenų užduotis, ugdymo tyrinys, įsisavinimo progresas, vidurinė mokykla

Received: 15 09 2010
Accepted: 05 10 2010