Evaluating Elementary Students' Response to Intervention in Written Expression

Sterett Mercer, Ioanna Tsiriotakis, Eun Young Kwon, Joanna Cannon

University of British Columbia

Paper Presented at the Canadian Society for the Study of Education (2019)

The purpose of this presentation is to (a) introduce data-based individualization (DBI) as a service delivery framework to improve outcomes for students with or at risk of learning disabilities, (b) introduce curriculum-based measurement (CBM) as an assessment approach for generating data to inform DBI, and (c) present an overview of findings identifying challenges and potential solutions for CBM in written expression.

Some academic interventions are generally more effective than others, as documented in the research literature; however, academic interventions that are highly effective on average are often not effective for specific students (Fuchs, McMaster, Fuchs, & Al Otaiba, 2013). This disconnect is particularly problematic given that special educators and related professionals are responsible for improving the academic skills of individual students. To address this concern, DBI is gaining recognition as an approach to special education service delivery (Fuchs et al., 2013). In DBI, generally effective interventions, as determined by the research literature, are first implemented, with ongoing monitoring of student academic outcomes through CBM. When CBM data indicate that the intervention is not effective for a specific student, the special educator engages in experimental teaching until CBM data indicate improvement. The DBI process of monitoring and modification is intended to be ongoing during service delivery

The ability to implement DBI rests on the availability of reliable and valid CBMs of the targeted academic skill. For reading, oral passage reading (i.e., having a student read

equivalent-difficulty passages while scoring for the number of words read correctly in one minute) performs well as a CBM reflecting overall reading skill and is sensitive to student skill improvements during intervention (Tindal, 2013). By contrast, efforts to develop a reliable and valid CBM in written expression for DBI have been problematic for two key reasons.

The first key difficulty in developing written expression CBM is that typical administration procedures (e.g., writing sample duration) do not yield reliable data for decision making. Because CBMs need to be efficient (i.e., quick to administer and score) to facilitate ongoing progress monitoring, CBM in written expression was originally developed with a very short writing duration (i.e., one minute to plan and three minutes to write) and with only one writing sample collected per occasion; however, research indicates that longer assessment durations and/or more than one writing sample per student are needed for reliable scores and to more closely approximate typical classroom writing assignments (McMaster & Espin, 2007).

In prior work (Keller-Margulis, Mercer, & Thomas, 2016), we investigated the reliability of seven minute writing samples, based on three narrative writing prompts per occasion, collected three times during one academic year for 145 students in grades 2-5. Generalizability theory analyses (Shavelson & Webb, 1991) were used to determine the effects of writing sample duration and the number of writing samples per occasion on the reliability of scores for screening and progress monitoring. In all grades, adequate reliability ($\geq$ .80) for absolute decisions about student performance could not be obtained with single writing samples of up to seven-minute duration, and reliability was not adequate for absolute decisions about student skill growth across the academic year even with three, seven-minute writing samples collected per assessment occasion. Our findings are consistent with the work of other research

teams finding that more than one longer duration writing sample, and samples from multiple writing genres, may be needed for reliable estimation of student writing skill (Graham, Hebert, Sandbank, & Harris, 2016; Kim, Schatschneider, Wanzek, Gatlin, & Al Otaiba, 2017).

The second key difficulty in developing written expression CBM is related to scoring complexity and feasibility. As noted above, more extensive writing samples (i.e., longer duration and multiple samples) are needed for reliable estimation of student writing skill, and increasing the length and number of student writing samples also increases the time required for teachers and other professionals to score writing samples, particularly for large numbers of students as is done in universal screening. This challenge is further compounded given that typical written expression CBM scoring methods have become more complex over time in an effort to improve the face and external validity of CBM scores. Early work in written expression CBM focused on simple scoring methods such as counts of the total number of words written (TWW); in addition to being perceived as poor indicators of overall writing quality by teachers (Ritchey & Coker, 2013), simple metrics like TWW correlate less strongly with other indicators of writing quality as student grade level increases and writing becomes more complex (McMaster & Espin, 2007). More complex scoring metrics such as correct word sequences (CWS; Videen, Marston, & Deno, 1982), i.e., counts of the number of adjacent words that are spelled correctly and make sense in context, have improved external validity compared to TWW for students in higher grades (e.g., Espin et al., 2008). CWS captures aspects of spelling, punctuation, syntax, and semantics, thereby more fully representing writing quality than TWW, but also considerably adding to scoring complexity and potentially affecting inter-scorer reliability.

In a preliminary attempt to resolve this tension between the need for scoring complexity to better represent writing quality with the time and effort required to score samples, we investigated the performance of automated text evaluation as a potential solution for scoring written expression CBM writing samples (Mercer, Keller-Margulis, Faith, Reid, & Ochs, 2019). Specifically, we investigated the ability of metrics generated by Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014), a program designed to evaluate characteristics of words, syntax, and discourse to predict student reading comprehension, to predict holistic ratings of writing quality for the writing samples from the general education sample in grades 2-5 described above (Keller-Margulis et al., 2016). In general, composite scores generated from Coh-Metrix output performed better than typical written expression CBM scores when predicting quality scores on the same narrative samples, and performed similarly to typical written expression CBM scores when predicting performance on other narrative writing samples. Although these results are encouraging for the potential for automated text evaluation to be useful in DBI, key limitations of this study are that (a) there were few students with disabilities in the sample, (b) writing sample duration (seven minutes) was below recommendation durations for reliable estimation of student skill, and (c) validity was evaluated in relation to other screening writing samples, rather than with standardized writing assessments.

The purpose of the current study is to address these limitations by investigating the performance of automated text evaluation to analyze the writing quality of 10-minute narrative writing samples produced by a sample of youth with significant learning difficulties or disabilities. The automated text evaluation quality scores are compared in relation to scores on

a standardized writing assessment to assess convergent validity, and in relation to scores on standardized reading and math assessments to assess discriminant validity. Change in the automated quality scores from the fall to spring of one academic year is also evaluated.

Method

Writing samples from students receiving one-on-one academic intervention through the Learning Disabilities Society of Greater Vancouver (LDS) were analyzed in this study. As part of intervention services, students served at LDS complete 10-minute picture prompted narrative writing samples at the beginning (Sep. - Oct.) and end (May - June) of each academic year to inform instructional goals and to monitor progress. The writing samples (n = 204) from 105 students in grades 2-12 were used to develop automated text evaluation models predicting writing quality in this study. We do not have detailed demographic or disability status information for these students; however, all students were experiencing academic difficulties substantial enough for parents or guardians to seek academic intervention services through a community agency (i.e., LDS).

Of the 105 students, a non-random sample of 33 were administered a standardized writing assessment in May - June of the same academic year, and also had scores on standardized assessments of reading and math (also collected in May - June). Of the 33 students, 7 were in grade 3, 5 in grade 4, 6 in grade 5, 5 in grade 6, 4 in grade 7, 4 in grade 8, and 2 in grade 9; 17 (51.5%) were male, and 31 (39.9%) had a learning goal in written expression at LDS (all 33 had learning goals in literacy).

Measures

Writing sample quality. Holistic writing quality for the picture-prompted writing samples was evaluated using the method of paired comparisons (Thurstone, 1927). Specifically, two raters on the research team each completed 3,000 comparisons of pairs of the 204 samples. In these comparisons, the writing sample that was of better overall quality was identified. These comparisons were tend processed by an algorithm to yield a writing quality score for each sample, i.e., higher scores indicated a greater likelihood that the sample would be rated better than other writing samples. Technical details of this process are presented in Mercer et al. (2019). Quality scores were highly consistent across the two raters, with r = .95.

Automated text evaluation. The writing samples were submitted to ReaderBench (Dascalu, Dessus, Trausan-Matu, Bianco, & Nardy, 2013), an open source program that provides many different indicators of vocabulary complexity and diversity, syntactic complexity, and text cohesion. The ReaderBench generated scores were then used to predict the rated sample quality scores using multiple machine learning algorithms (see Mercer et al., 2019, for procedural details). Of the algorithms explored, partial least squares (PLS) regression was the best performing (85% of the variance in quality ratings explained), and presented results are based on the model-predicted quality scores from the PLS algorithm that are based on automated text evaluation with ReaderBench.

Standardized writing assessment. Two subtests of the Test of Written Language, 4th edition (Hammill & Larsen, 2009) were scored based on one picture-prompted narrative sample (5 minutes to plan, 15 minutes to write). Scoring for Contextual Conventions considers spelling and grammatical errors, and scoring for Story Composition considers quality of vocabulary, plot,

and interest to the reader. Raw scores, rather than age- or grade-based standard scores, were analyzed given that the sample was quite heterogenous in age/grade.

Standardized reading and math assessments. Computerized adaptive assessments of broad reading and broad math skills were administered (aReading and aMath; Christ et al., 2014). Each assessment requires approximately 20 minutes to administer and is designed to assess skill levels ranging from kindergarten to grade 12 on a comparable scale. More detailed reliability and validity information is available at https://charts.intensiveintervention.org/chart/academic-screening

Results

We first examined the extent to which the automated writing quality composite scores based on the fall and spring picture-prompted writing samples correlated with spring scores on the standardized writing, reading, and math assessments (see Table 1).

Table 1. *Automated quality scores in relation to standardized writing, reading, and math scores*

	TOWL CC	TOWL SC	aReading	aMath
	r (R^2)	*r* (R^2)	*r* (R^2)	*r* (R^2)
Fall Quality	.69 (.48)	.47 (.22)	.53 (.28)	*.24 (.06)*
Spring Quality	.76 (.57)	.53 (.28)	.56 (.31)	*.35 (.12)*
TOWL Quality	.78 (.60)	.69 (.48)	--	--

Note. n = 33. TOWL = Test of Written Language (4th ed.), CC = Contextual Conventions, SC = Story Composition. Values in italics are not statistically significant (α = .05).

Several key findings are evident in the values in Table 1. First, there was a general pattern of stronger correlations of automated writing quality scores in relation to writing compared to reading and math scores, which were not statistically significant, providing both convergent and discriminant validity evidence. Second, the automated quality scores had stronger relations to the TOWL subtest assessing mechanics and conventions than the TOWL subtest assessing substantive writing quality. Third, there is evidence of the generalizability of the automated scoring model—the final row of the table presents correlations of automated writing quality for the TOWL writing sample, based on the model developed for the other picture prompted samples, in relation to the actual TOWL scores. Last, although the full results are not presented in the table, the scores from automated text evaluation demonstrated either comparable or improved validity compared to typical written expression CBM hand scoring methods. For example, counts of the Total number of Words Written (TWW) correlated with TOWL CC at $r = .47$ for the fall sample and $r = .59$ for the spring sample. More complex scoring with Correct Word Sequences (CWS) correlated with TOWL CC at $r = .67$ on both the fall and spring samples.

After exploring the validity of the automated quality scores, we did preliminary analyses to determine the extent to which the scores demonstrated growth from the fall to the spring. Overall, there was a statistically significant improvement ($p < .001$) in automated quality scores from the fall ($M = -.05$) to the spring ($M = .17$). Although limited by the very small number of students in each grade, there was a general pattern of higher scores as grade level increased, and students in all grades showed improvements from fall to spring (see Figure 1).

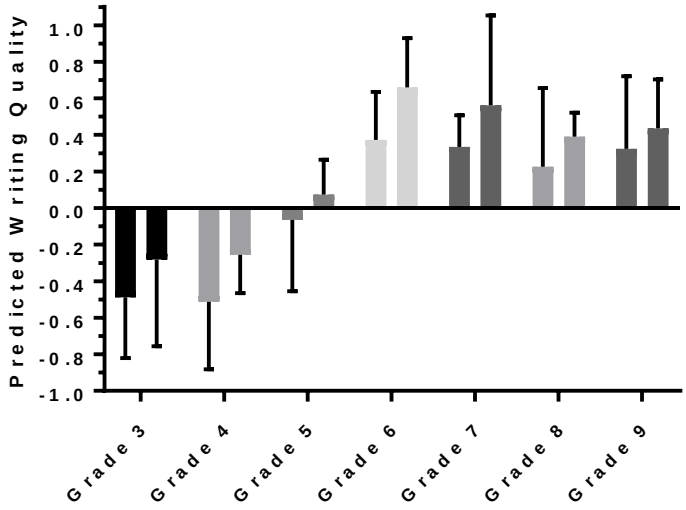

Figure 1. Automated writing quality scores by grade in fall and spring (*n* = 33).

Discussion

These results provide preliminary validity evidence that automated writing evaluation can be used to score picture-prompted writing samples to monitor the writing skills of students with significant learning difficulties. These findings are of particular importance given that scoring with automated text evaluation potentially can substantially improve both the scoring feasibility of written expression CBM and the validity of written expression CBM as a screening and progress monitoring tool. By reducing the time required for teachers to evaluate overall writing quality in samples, it is possible that using automated text evaluation for that purpose can enable teachers to provide more frequent and detailed feedback on specific aspects of student writing in need of improvement (Wilson & Czik, 2016); importantly, this work on automated text evaluation is not intended to imply that computer scoring can replace teacher evaluation of student writing, rather automated text evaluation may be a useful complement to the more detailed teacher evaluation that is essential for improving student writing skill.

Providing effective special education services to students requires educators to make good decisions about whether and when to modify instruction to improve student outcomes, and these decisions can only be made when we have good data on student academic skill growth. Unlike reading, developing such a measure in written expression has been challenging; however, these results provide preliminary information that may support the development of an automated tool to help teachers and other school staff identify students in need of intervention (relative to local norms) and to monitor their response to targeted instruction and intervention.

Acknowledgements

This work was supported by funding from the Social Sciences and Humanities Research Council of Canada (SSHRC), the U.S. Department of Education (Institute of Education Sciences), and the Chris Spencer Foundation. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Christ, T. J., Arañas, Y. A., Kember, J. M., Kiss, A. J., McCarthy-Trentman, A., Monaghen, B. D., . . . White, M. J. (2014). *Formative Assessment System for Teachers Technical Manual: earlyReading, CBMReading, aReading, aMath, and earlyMath.* Minneapolis, MN: Formative Assessment System for Teachers.

Dascalu, M., Dessus, P., Trausan-Matu, Ş., Bianco, M., & Nardy, A. (2013). ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education: 16th International Conference Proceedings* (pp. 379-388). Berlin, DE: Springer.

Espin, C. A., Wallace, T., Campbell, H., Lembke, E. S., Long, J. D., & Ticha, R. (2008). Curriculum-based measurement in writing: Predicting the success of high-school students on state standards tests. *Exceptional Children, 74*, 174-193. doi:10.1177/001440290807400203

Fuchs, D., McMaster, K. L., Fuchs, L. S., & Al Otaiba, S. (2013). Data-based individualization as a means of providing intensive instruction to students with serious learning disorders. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (2nd ed., pp. 526-544). New York: Guilford.

Graham, S., Hebert, M., Sandbank, M. P., & Harris, K. R. (2016). Assessing the writing achievement of young struggling writers: Application of Generalizability Theory. *Learning Disability Quarterly, 39*(2), 72-82. doi:10.1177/0731948714555019

Hammill, D. D., & Larsen, S. C. (2009). *Test of written language 4 (TOWL-4)*: San Antonio, TX: Pro-Ed Assessments.

Keller-Margulis, M. A., Mercer, S. H., & Thomas, E. L. (2016). Generalizability theory reliability of written expression curriculum-based measurement in universal screening. *School Psychology Quarterly, 31*(3), 383-392. doi:10.1037/spq0000126

Kim, Y.-S. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and Writing, 30*(6), 1287-1310. doi:10.1007/s11145-017-9724-6

McMaster, K. L., & Espin, C. A. (2007). Technical features of curriculum-based measurement in writing. *Journal of Special Education, 41*(2), 68-84. doi:10.1177/00224669070410020301

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

Mercer, S. H., Keller-Margulis, M. A., Faith, E. L., Reid, E. K., & Ochs, S. (2019). The potential for automated text evaluation to improve the technical adequacy of written expression curriculum-based measurement. *Learning Disability Quarterly, 42*(2), 117-128. doi:10.1177/0731948718803296

Ritchey, K. D., & Coker, D. L., Jr. (2013). An investigation of the validity and utility of two curriculum-based measurement writing tasks. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 29*(1), 89-119. doi:10.1080/10573569.2013.741957

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273-286. doi:10.1037/h0070288

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education, 2013*, 1-29. doi:10.1155/2013/958530

Videen, J., Marston, D., & Deno, S. (1982). Correct word sequences: A valid indicator of proficiency in written expression (Vol. IRLD-RR-84, pp. 61). *Minnesota Univ, Minneapolis Inst for Research on Learning Disabilities*.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94-109. doi:10.1016/j.compedu.2016.05.004