Journal of School Psychology 68 (2018) 19–37

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4

Joshua Wilson*

University of Delaware, United States

ARTICLE INFO

Action Editor: Milena Keller-Margulis

Keywords:
Writing
Screening
Automated essay scoring
Project Essay Grade (PEG)
Common Core State Standards
Smarter Balanced assessment consortium
Literacy
ROC curve analysis

ABSTRACT

The adoption of the Common Core State Standards and its associated assessments has placed increased focus on writing performance. Consequently, weak writers may be at risk of failing Common Core English language arts (ELA) assessments. Thus, the current study sampled a diverse group of third and fourth grade students (n = 100 and 130, respectively) who were administered Fall and Spring writing screeners using the procedures of a Direct Assessment of Writing (DAW). Results were used to predict whether students did or did not attain grade-level standards as evaluated by the summative Smarter Balanced ELA assessment. Writing screeners were scored using the Project Essay Grade (PEG) automated essay scoring system. ROC curve analysis and logistic regression were used to evaluate screening models. Area under the ROC curve (AUC) values for grade 3 were in the acceptable range (Fall = 0.74, Spring = 0.75). AUCs approached or fell within the excellent range for grade 4 (Fall = 0.79, Spring = 0.83). Sensitivity-based and d-based cutpoints were selected and measures of diagnostic accuracy, including sensitivity and specificity, are reported. Results indicate that automatically-scored DAW has promise for universal screening for writing risk.

In the United States, 42 states, the District of Columbia, and four territories have adopted the Common Core State Standards (CCSS) and associated English Language Arts (ELA) assessments. These standards and assessments establish rigorous benchmarks for both reading and writing achievement, but their focus on writing achievement represents a dramatic increase over prior literacy reform initiatives (Graham & Harris, 2013; Shanahan, 2015). Indeed, Common Core-aligned ELA assessments include several short and extended constructed-response items as well as writing performance tasks (Behizadeh & Pang, 2016).

Scores from Common Core ELA assessments are used to index the degree to which students have attained grade-level Common Core reading and writing standards. While it is left to individual states and school districts to determine the types of decisions attached to scores from these assessments, historically poor performance on summative assessments may result in referring the student for intervention and remediation (Graham, Hebert, & Harris, 2011; Jones et al., 1999); tracking the student into specific classes, programs, or schools (Decker & Bolt, 2008; Goertz & Duffy, 2003); or retaining the student from advancing to the next grade (Darling-Hammond, 2004; Hamilton et al., 2007). In addition, students who fail large-scale accountability assessments are at increased risk of referral to special education and school dropout (Figlio & Getzler, 2002; Heubert & Hauser, 1999).

Given such consequences, educators are increasingly adopting prevention/intervention frameworks, such as Response to Intervention (RTI; Fuchs, Mock, Morgan, & Young, 2003), to ensure the accurate and timely identification and remediation of students at risk of failing to achieve grade-level standards. A core component of these frameworks is universal screening, a process designed to efficiently and accurately identify students at risk of experiencing academic or behavioral difficulties (Fuchs et al., 2003;

* University of Delaware, School of Education, 213E Willard Hall Education Building, Newark, DE 19716, United States.
E-mail address: joshwils@udel.edu.

https://doi.org/10.1016/j.jsp.2017.12.005
Received 13 January 2017; Received in revised form 21 October 2017; Accepted 18 December 2017

Sugai & Horner, 2002).

Accordingly, the present study considered the role of universal screening for writing difficulties within the context of Common Core. Given the focus on writing achievement in the CCSS, the degree to which writing skills are assessed within Common Core ELA assessments, and well documented associations between reading and writing skills (Abbott, Berninger, & Fayol, 2010; Ahmed & Wagner, 2014; Fitzgerald & Shanahan, 2000; Graham & Hebert, 2010), it is reasonable that writing screeners may accurately identify students at risk of failing Common Core ELA assessments. Indeed, writing skills have predicted unique variance on tests of reading, particularly if those tests include constructed-response items (Jenkins, Johnson, & Hileman, 2004; Johnson, Jenkins, & Jewell, 2005).

Specifically, the present study considered whether a direct assessment of writing (DAW) scored using automated essay scoring (AES) can accurately screen and identify third and fourth graders at risk of failing the Smarter Balanced ELA test. Despite the abundance of research showing that AES yields reliable scores (see Shermis, 2014; Shermis & Burstein, 2003, 2013), very little research has explored validity evidence related to the use of AES for specific purposes (Klobucar, Elliot, Deess, Rudniy, & Joshi, 2013), such as universal screening. Thus, the current study also was intended to contribute to incipient research that evaluates validity evidence pertaining to the use of AES within K-12 prevention/intervention frameworks.

1. Universal screening for writing

Research on writing screening has primarily focused on the use of curriculum-based measurement for writing (W-CBM), brief writing tasks administered and scored according to standardized procedures (Deno, Marston, & Mirkin, 1982; Espin et al., 2000; Parker, Tindal, & Hasbrouck, 1991). W-CBM involves providing students with a stimulus, either a picture-word prompt or sentence-writing prompt for early elementary-grade students (Coker & Ritchey, 2010; McMaster, Du, et al., 2011), or narrative or expository prompts for students at grades three or higher (Espin et al., 2000; Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006). Students are given 1 min to think and between 3 and 10 min to compose a response—administration times vary according to students' grade level and publisher guidelines (Hosp, Hosp, & Howell, 2007). Responses are scored using general outcome measures of writing proficiency such as total words written (TWW; a measure of transcription fluency), words spelled correctly (WSC; a measure of spelling accuracy), and correct word sequences (CWS; a measure of transcription fluency accounting for accurate spelling, grammar, semantics, capitalization, and punctuation).

Scores from W-CBM generally show strong evidence of reliability (i.e., inter-rater reliability, alternate-form reliability) and moderate evidence of criterion validity with district and state ELA accountability assessments (Espin et al., 2000; Furey, Marcotte, Hintze, & Shackett, 2016; Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002; McMaster & Campbell, 2008; Mercer, Martinez, Faust, & Mitchell, 2012). For instance, a recent meta-analysis reported that, when averaging across W-CBM metrics and grade levels, the mean correlation between W-CBM and state-developed tests was 0.61 [0.51, 0.69] (Romig, Therrien, & Lloyd, 2016).

However, documenting evidence of reliability and criterion validity via correlation is not equivalent to establishing that scores from W-CBM support accurate screening decisions (Wilson & Reichmuth, 1985). This requires a different analytic procedure such as Receiver Operating Characteristic (ROC) curve analysis or logistic regression that directly calculates the degree to which students are correctly classified as at risk and not at risk.

Three prior studies examined classification accuracy of W-CBM for screening for risk of failing a state writing or ELA assessments. Lopez and Thompson (2011) examined classification accuracy of 3-minute story starter probes administered to students in grade 6 (n = 36), grade 7 (n = 23), and grade 8 (n = 24) and scored for CWS. A cutpoint of 1 standard deviation below the mean was selected to predict students who scored in the "Does not meet expectations" range of Arizona's state writing test. Corresponding base rates for the three grades were 17%, 22%, and 4%. The authors did not use ROC curve analysis, but calculated overall classification accuracy via a confusion matrix. Correct classification rates were 75%, 87%, and 96%. However, since the base rates were quite low, these values are misleading and may overestimate the utility of CWS as a screener. With these base rate values, correct classification rates of 83%, 78%, and 96% would have been achieved simply by assuming that no students were at risk. A measure should significantly improve classification accuracy over what would be achieved simply by identifying no at risk students (Johnson, Jenkins, Petscher, & Catts, 2009; Meehl & Rosen, 1955).

Keller-Margulis, Payan, Jaspers, and Brewton (2016) examined classification accuracy of several W-CBM metrics applied to 3-minute story starter probes administered in the Fall, Winter, and Spring of fourth grade. Samples were scored for production-dependent measures [TWW, WSC, CWS, correct punctuation (CP)], production-independent measures (%WSC, %CWS), and accuracy of production (correct minus incorrect word sequences [CIWS]). Using ROC curve analyses, the authors sought to identify cutscores that yielded sensitivity and specificity values ≥ 0.70 when predicting lack of proficiency on the Texas state writing test. Sensitivity is the true positive rate, and specificity is the true negative rate. Separate ROC curve analyses were conducted with samples of native English speakers, English Language Learners (ELLs), and exited ELLs who were monitored for academic success. Sample sizes were small for each group (n = 89, 18, and 30), as were the base rates in each sample (23.6%, 21.1%, and 3.2%). Results of the study indicated that production-dependent measures, which are frequently relied upon in W-CBM, did not yield acceptable classification accuracy. Instead, the production-independent measures did better, as did the accuracy-of-production measure (CIWS). No measure was consistently adequate across all time points and for each of the three samples; however, %CWS was closest, with acceptable classification at all three time points for the native English speakers, at the Winter benchmark for ELLs, and in the Spring for monitored students.

Finally, a study by Furey et al. (2016) found similar results in a sample of fourth graders whose performance on the written expression subtest of the Massachusetts state assessment was predicted from 3-minute and 10-minute W-CBM story prompts. Prompts were scored for TWW, CWS, %CWS, and CIWS. The base rate in this sample was 55%. ROC curve analyses were used to examine