Classification Accuracy and Efficiency of Writing Screening Using Automated Essay Scoring

Joshua Wilson[a]

University of Delaware

Jessica Rodrigues[b]

University of Missouri

Author Note

[a]Joshua Wilson, School of Education, University of Delaware; [b]Jessica Rodrigues, Department of Education, University of Missouri.

Correspondence concerning this article should be addressed to Joshua Wilson, Ph.D., University of Delaware, School of Education, 213E Willard Hall Education Building, Newark, DE, 19716, United States. Tel: +13028312955. Email: joshwils@udel.edu.

Acknowledgements

This research was supported in part by Delegated Authority contract EDUC432914160001 from Measurement Incorporated® and by Grant R305H170046 from the Institute of Education Sciences, U.S. Department of Education, to the University of Delaware. The opinions expressed in this paper are those of the authors and do not represent the views of Measurement Incorporated, the Institute, or the U.S. Department of Education, and no official endorsement by these agencies should be inferred.

Suggested Citation:

Wilson, J., & Rodrigues, J. (2020). Classification accuracy and efficiency of writing screening using automated essay scoring. *Journal of School Psychology*, *82*, 123-140.

Abstract

The present study leveraged advances in automated essay scoring (AES) technology to explore a proof of concept for a writing screener using the *Project Essay Grade* (PEG) program. First, the study investigated the extent to which an AES-scored multi-prompt writing screener accurately classified students as at risk of failing a Common Core-aligned English language arts state test. Second, the study explored whether a similar level of classification accuracy could be achieved with a more efficient form of the AES-screener with fewer writing prompts. Third, the classification accuracy of the AES-scored screeners was compared to that of screeners scored for word count. Students in Grades 3-5 (n = 185, 167, and 187, respectively) composed six essays in response to multiple writing-prompt screeners on six different randomly assigned topics, two essays in each of three different genres (narrative, informative, and persuasive). Receiver operating characteristic (ROC) curve analysis was used to assess classification accuracy and to identify multiple cut scores with associated sensitivity and specificity values, and positive and negative posttest probabilities. Results indicated that the AES-scored multi-prompt screener and screeners with fewer prompts yield acceptable classification accuracy, are efficient, and are more accurate than screeners scored for word count. Overall, results illustrate the viability of writing screening using AES.

Keywords: Writing; Screening; Automated essay scoring; At risk; Writing assessment.

Classification Accuracy and Efficiency of Writing Screening Using Automated Essay Scoring

In the United States, there is increased attention on improving students' writing skills. This is reflected in the focus on writing in the Common Core (National Governors Association and Council of Chief State School Officers, 2010) and other contemporary state standards, as well as standards-aligned assessments of English language Arts (ELA) that include writing performance assessments (Behizadeh & Pang, 2016). Thus, students who struggle with writing are at increased risk of failing state ELA assessments, as well as experiencing a number of other difficulties in K-12 and post-secondary settings (Graham & Hebert, 2010; National Commission on Writing for America's Families, Schools, and Colleges [NCWAFSC], 2004, 2005).

In order to identify struggling writers before they experience these negative outcomes, researchers and educators are investigating whether writing assessments delivered early in the school year as part of a universal screening process can accurately identify students at risk for failing state ELA assessments (Espin et al., 2008; Furey, Marcotte, Hintze, & Shackett, 2016; Keller-Margulis, Payan, Jaspers, & Brewton, 2016; Wilson, 2018). Universal screening involves assessing all students in a school. Thus, assessments used for universal screening tend to be formative general outcome measures that quickly index a student's overall proficiency as opposed to diagnostic measures that are reserved for a small portion of the school population who have been identified for intervention. Key criteria for selecting a measure for use in universal screening are (a) accuracy in classifying students as at risk or not at risk (i.e., true positives and true negatives) and (b) efficiency of administration and scoring (Jenkins, Hudson, & Johnson, 2007). The former attribute is known as classification accuracy, which can be defined in multiple ways. For the purposes of the present study, it is defined as the extent to which classifications based on

an observed cut score correspond with actual classifications using a validated outcome measure (Lee, 2010).

The present study extended existing research on writing screeners for use in universal screening by exploring a proof of concept for a writing screener that used automated essay scoring (AES). AES is a form of computer-based scoring that relies on natural language processing tools to identify text features associated with writing quality and that utilizes an algorithm to weight and combine those features to produce writing-quality scores (see Shermis & Burstein, 2003, 2013). The efficiency and consistency of AES are desirable characteristics for a screening measure: AES is a 100% consistent scoring method that provides immediate scoring capabilities without the need for human scoring. However, little is known about whether AES scores can accurately classify students into risk categories or whether AES scores more accurately classify students than easily obtainable measures such as word count. Thus, the present study assessed the classification accuracy and efficiency of a writing screener scored using an AES system called *Project Essay Grade* (PEG; Page, 2003; Shermis, Koch, Page, Keith, & Harrington, 2002). In addition, the study compared the classification accuracy of scores from PEG to that of word count.

Challenges Facing Writing Screening

High base rate. There are a number of challenges to writing screening. First among them is the likelihood that writing screening will be conducted in the presence of a high base rate—a high base rate indicates that a given condition is highly prevalent within a population. Nationally, many students fail to achieve writing proficiency (Persky, Daane, & Jin, 2002) and Tier 1 elementary writing instruction is often limited, with infrequent use of evidence-based practices (Brindle, Graham, Harris, & Hebert, 2016; Gilbert & Graham, 2010). Within an MTSS framework, the presence of a high base rate signals the need to improve the quality of Tier 1 writing instruction,

a need that is well recognized (e.g., Graham, Bollinger, et al., 2012; Graham, McKeown, Kiuhara, & Harris, 2012).

However, the presence of high base rates does not itself negate the need to identify accurate and efficient writing screeners. Indeed, when students' probability of risk is between 10-50%, scholars recommend administering assessments to more accurately classify students into risk categories (VanDerHeyden, 2013). Rather, the presence of high base rates necessitates that the use of the screener does not detract from efforts to improve the quality of Tier 1 writing instruction. One way to ensure screening does not detract from such efforts would be to leverage a highly recommended Tier 1 instructional method for the purpose of screening, such as having students practice writing, especially the kind of writing that requires students to build stamina and to plan, draft, revise, and edit (Graham, Bollinger, et al., 2012; Graham, McKeown, et al., 2012). As will subsequently be discussed, the kind of writing elicited by existing curriculum-based writing screeners does not support this kind of writing practice; hence, we explore an alternative screening method.

Measurement error. Another issue facing writing screening, and writing assessment generally, pertains to measurement error, which is variance in observed scores not attributed to the object of measurement (i.e., students; see Bouwer, Béguin, Sanders, & van den Bergh, 2015; Kim, Schatschneider, Wanzek, Gatlin, & Al Otaiba, 2017). A principal source of measurement error in writing screener scores is rater error, arising from inconsistent interpretations and applications of scoring criteria (Godshalk, Swineford, & Coffman, 1966; Wolfe, 2006). Other sources of measurement error pertain to aspects of the writing task, such as the genre and the topic of the prompt. The genre refers to the rhetorical purpose (e.g., to inform, to persuade, to entertain) and structure (e.g., an essay, an argument, a story) that students are expected to fulfill and adopt when

responding to a writing prompt. The prompt topic refers to the specific content that students are to write about (e.g., outer space, a favorite animal, school on Saturday). Prior research has shown that writing performance varies by genre (Davidson & Berninger, 2016; Kim et al., 2017; Mercer, Martinez, Faust, & Mitchell, 2012; Olinghouse & Wilson, 2013) and by prompt topics (Keller-Margulis, Mercer, & Thomas, 2016), likely because different writing tasks call upon different genre knowledge and content knowledge (Bereiter & Scardamalia, 1987; Olinghouse & Graham, 2009; Olinghouse, Graham, & Gillespie, 2015).

To the degree that raters, genres, and topics contribute measurement error to a writing screener, scores from that screener will have diminished reliability and, consequently, those scores are unlikely to support accurate classification decisions. The presence of measurement error across measurement facets requires one of two solutions. One solution is to obtain a greater number of observations across each facet – such as using multiple raters or obtaining multiple prompts written in multiple genres about multiple topics – and average the scores to minimize error (Brennan, 2001; Shavelson & Webb, 1991). Of course, in the context of universal screening, this solution may be infeasible because sampling greater numbers of raters or writing prompts reduces efficiency. A second solution is to minimize error variance across those facets, for instance by employing rigorous rater training methods or utilizing scoring methods that are more consistent and less sensitive to genre and topic variance. As will subsequently be described, the present study explores a screening method that draws on both of these solutions.

Challenges of Curriculum-Based Measurement for Writing Screening

The most researched method of writing screening is curriculum-based measurement of writing (CBM-W). CBM-W involves providing students a brief amount of time to respond to a prompt – typically 1min to plan and 3min to write – and scoring the resultant text for one or more

low-inference general outcome measures of writing quality, such as counting the total words written or the number of words spelled correctly; calculating the number of correct word sequences absent of errors in spelling, grammar, and semantics; or calculating the difference between correct and incorrect word sequences (Deno, Marston, & Mirkin, 1982; Espin et al., 2008; Espin, Shin, Deno, Skare, Robinson, & Benner, 2000; Furey et al., 2016; Parker, Tindal, & Hasbrouck, 1991). Although CBM-W have demonstrated desirable psychometric properties in some contexts (Gansle et al., 2004; Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006; Romig, Therrien, & Lloyd, 2017), recent evidence calls into question the suitability of CBM-W for screening purposes (Furey et al., 2016; Keller-Margulis, Payan et al., 2016).

First, CBM-W screening methods may be insufficiently reliable to support accurate screening. Scores of CBM-W are susceptible to rater error (Allen, Poch, & Lembke, 2018) as well as prompt and genre effects because they are scored by humans and typically involve the use of single prompt assessments for making classification accuracy decisions (c.f., Furey et al., 2016). For these reasons, a single CBM-W screener is insufficiently reliable for use in universal screening. Keller-Margulis, Mercer, and Thomas (2016) revealed that reliable relative decisions (i.e., rank ordering) would require averaging scores from a minimum of three 3min prompts or two 5-7min prompts, and that reliable absolute decisions (i.e., at risk/not at risk) would require averaging across scores from a minimum of three 5-7min prompts. The authors lamented that adopting more reliable screening procedures would likely be infeasible because of the time and financial costs of scoring CBM-W with human raters.

Second, CBM-W scoring measures solely evaluate the constructs of fluency and accuracy. Measures evaluating the broader construct of writing quality are not employed – writing quality represents the degree to which a text achieves a rhetorical purpose and displays the characteristics

of "good" writing (Deane, 2013; Slomp, 2012). The construct of writing quality is typically measured by one or more raters who apply a rubric to generate a single score to represent overall writing quality or multiple scores to represent writing quality across a number of dimensions, such as development of ideas, organization, sentence fluency, and word choice. The former method is referred to as holistic scoring and the latter method is referred to as analytic scoring (Huot, 1990). Holistic scoring is commonly used to evaluate writing quality on large-scale writing assessments, such as the National Assessment of Educational Progress (NAEP; National Assessment Governing Board, 2017), and analytic scoring, particularly the Six Trait model (Northwest Regional Educational Laboratory, 2004), is widely-used to support instructional decision making (Coe, Hanita, Nishioka, & Smiley, 2011). Although measures of writing quality better capture the construct of writing ability than measures of fluency (Deane, 2013), and as such may support more accurate screening decisions (Wilson, Olinghouse, McCoach, Andrada, & Santangelo, 2016), measures of writing quality are not employed in CBM-W because they are more time-consuming and subjective to score than measures of fluency. However, advances in automated scoring afford the possibility of utilizing measures of writing quality in assessments used for universal screening.

Using Automated Scoring to Advance Writing Screening

The present study examined writing screening using automated essay scoring (AES) software, an increasingly prevalent form of educational technology in U.S. schools (Stevenson & Phakiti, 2014; Wilson & Czik, 2016; Wilson & Roscoe, 2020). AES and automated feedback systems are developed using natural language processing (NLP) techniques that identify language/text features that measure, or serve as proxies for, key dimensions of writing quality, such as development of ideas, organization, style, sentence structure, vocabulary, and conventions (Deane, 2013). NLP features are defined a priori or induced from the data via regression or

machine-learning methods that weight and combine features to maximize prediction of human ratings of writing quality. Key benefits of AES are efficiency – scores are returned immediately – and reliability and consistency: AES has been shown to be as reliable (or more so) than human raters (Shermis, 2014) and AES is 100% consistent.

Although formative AES systems are regarded positively (see Palermo & Thomson, 2018; Roscoe & McNamara, 2013; Wilson & Roscoe, 2020), there are concerns regarding the construct validity of AES due to high correlations with text length (Deane, 2013; Perelman, 2014) and poor discrimination across different dimensions of writing quality (Wilson, Olinghouse, & Andrada, 2014). That said, these problems are not unique to AES. Human ratings of writing quality also evince high correlations with text length (Perelman, 2012) and human raters using the Six Trait scoring model fail to discriminate across dimensions of writing quality (Gansle et al., 2006). Thus, while the efficiency and reliability of AES address two major barriers facing the use of measures of writing quality for screening, open questions remain regarding the validity of AES for use in universal screening.

Two recent studies illustrate the promise of AES for supporting writing screening. The first study, a generalizability study of the Project Essay Grade (PEG) system in Grades 3-5 showed that reliable low-stakes decisions (G and $\phi \geq .80$) could be made about non-struggling writers by averaging scores from three randomly assigned writing prompts, one prompt per genre, and that reliable low-stakes decisions about struggling writers could be made by averaging scores from six writing prompts, two prompts per genre (Wilson, Chen, Sandbank, & Hebert, 2019). These findings were promising, given that prior generalizability studies of human-scored writing assessments identified the need for far greater number of scores (in the range of 12-16) to make reliable decisions (Graham, Hebert, Sandbank, & Harris, 2016; Keller-Margulis, Mercer, &

Thomas, 2016; Kim et al., 2017; Schoonen, 2005, 2012). The findings of Wilson et al. (2019) suggested that PEG may afford a feasible method of reliably assessing students' general writing proficiency for the purpose of universal screening.

A second study directly assessed the potential of AES for use as a universal screener. Wilson (2018) utilized receiver operating characteristic (ROC) curve analysis to evaluate the overall classification accuracy of cut scores derived from a single writing prompt screener scored by PEG and administered to third- and fourth-graders at two time points (Fall and Spring) for classifying students who passed or failed a Common Core-aligned state English language arts (ELA) test called Smarter Balanced. Overall classification accuracy of the PEG-derived cut scores was in the acceptable range, with area under the curve (AUC) values falling between .74 (Fall Grade 3) and .83 (Spring Grade 4). Additionally, cut scores were selected that maximized sensitivity (true positive rate) and specificity (true negative rate) in a manner consistent with recommendations in the field (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009; Smolkowski & Cummings, 2015). Although Wilson (2018) assessed writing in only one genre (argumentative writing) and with only one topic per occasion, results showed promise of AES for use as a universal screener.

The Present Study

The present study was conducted to address some of the measurement challenges and limitations facing existing writing screening and to further explore a proof of concept for the use of AES for universal screening. The study is described as a proof of concept because, as detailed in the Methods, students responded to writing prompts by hand and did not keyboard their responses directly into the AES system. Capitalizing on the efficiency of AES necessitates the use

of technology to respond to prompts; hence, the present study does not afford a test of the intended implementation of AES for screening but does illustrate a proof of concept.

Given the findings of Wilson et al. (2019) that reliable low-stakes decisions could be made about struggling writers using two writing prompts per genre scored via PEG, students in Grades 3-5 in the present study were screened for risk of failing the Smarter Balanced ELA test based on the average PEG score from six writing prompts (2 prompt topics written in 3 different genres). By administering multiple writing prompts in multiple writing genres, we obtained a broader, more representative and reliable sample of students' writing ability than has been obtained in prior screening studies. Further, by using AES, a perfectly consistent scoring method, we maximized score reliability. Thus, our first research question was: (1) How accurate was the AES-scored multi-prompt writing screener for identifying students at risk of failing the Smarter Balanced ELA test?

Our decision to use the Smarter Balanced ELA test as the criterion measure reflected our recognition that writing proficiency is one of multiple proficiencies comprising a broader literacy construct (National Governors Association and Council of Chief State School Officers, 2010). Writing is related to these other literacy skills (i.e., reading and oral language) via shared linguistic resources (Berninger, 2000; Berninger & Abbott, 2010; Fitzgerald & Shanahan, 2000; Schoonen, 2019) and writing performance is correlated with performance on tests of other literacy skills (Abbott, Berninger, & Fayol, 2010; Ahmed, Wagner, & Lopez, 2014). Thus, we utilized a single-proficiency screener to predict performance on an assessment of multiple proficiencies, an approach to screening that has been used productively in the fields of reading and mathematics (e.g., Good, Simmons, & Kame'enui, 2001; Jordan, Glutting, Ramineni, & Watkins, 2010; Rodrigues, Jordan, & Hanson, 2019; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008).

In addition, given that (a) efficiency and feasibility are critical for writing screeners, and (b) the study by Wilson (2018) showed that acceptable classification accuracy was achieved with a single writing prompt scored by PEG in Grades 3 and 4, the present study also explored whether a similar level of classification accuracy could be achieved between the six-prompt writing screener and writing screeners that contained as few as one prompt. Therefore, our second research question was: (2) Is it possible to retain a similar level of classification accuracy by implementing an AES-scored writing screener with fewer than six writing prompts?

Finally, given research showing high correlations between AES and word count, our final research question was: (3) How does the classification accuracy of AES-scored writing screeners compare to writing screeners scored for word count? Given the costs associated with implementing AES, we considered whether an AES measure of writing quality supports more accurate screening than word count, an easily calculated, less expensive measure of writing fluency.

Methods

Participants

Participants in the present study included 539 students in Grade 3 (n = 185), Grade 4 (n = 167), and Grade 5 (n = 187). All participants were also included in a prior study by Wilson et al. (2019). Consistent with the project's IRB approval, parental consent and student assent were obtained. Students were sampled from 31 classrooms (Grade 3 = 13; Grade 4 = 9; Grade 5 = 9) in three elementary schools in a large and diverse mid-Atlantic school district. All students in Grades 3-5 were eligible for participation in the study. No students were removed, or their data deleted, as the present study employed multiple imputation procedures to handle missing data due to absences on the days when the writing screeners were administered and due to attrition when the Smarter Balanced ELA test was administered (see "Multiple Imputation" section, below).

Table 1 presents study demographics disaggregated by grade-level. The sample was diverse, including a sizeable proportion of Hispanic/Latinx students and Spanish-speaking English-language Learners (ELLs; >90% of ELLs), reflecting the demographics of the school district. Information on students' language proficiency was unavailable for reporting. However, the district's approach to ELL instruction involved immersion in English-speaking classes with supplemental small-group support by an ELL specialist. Although students' individual socio-economic status was unavailable, 36% of the district's students qualified as low-income.

Chi-square tests compared demographics across grades. Results indicated that there was no statistically significant difference across grades in the proportion of males and females (χ^2 = 3.01, df = 2, p = .222), students in different racial categories (χ^2 = 8.36, df = 6, p = .213), students with disabilities (χ^2 = 2.38, df = 2, p = .304), or ELLs (χ^2 = 0.71, df = 2, p = .700). Finally, as shown in Table 1, the grade-level samples demonstrated an average level and range of writing performance, as measured by age-based standard scores (M = 100, SD = 15) of the Sentence Composition subscale of the Wechsler Individual Achievement Test 3rd Edition (WIAT III).

Measures

Smarter Balanced ELA test. The Smarter Balanced ELA test is a computer-based Common-Core aligned assessment currently administered to students in Grades 3-8 and high school in 12 states, the U.S. Virgin Islands, and the Bureau of Indian Affairs. More than six million students take this summative test each year (see http://www.smarterbalanced.org/about/). The test evaluates Common Core reading, writing, listening, and research standards. It includes a combination of selected-response items delivered using a computer-adaptive testing (CAT) framework, and short and extended constructed response items (i.e., performance tasks). All CAT items on the Smarter Balanced ELA test are machine scored and assess differing depths of

knowledge within a domain; correct responses to CAT items with higher depth of knowledge result in higher test scores. The performance tasks are intended to be either hand-scored or machine scored. In all, the CAT portion of the ELA test is expected to take 1.5 hours and the performance task to take 2 hours (Smarter Balanced Assessment Consortium [SBAC], 2017a). The test was administered in Spring 2016 within the state-determined assessment window, mid-March to mid-May by district teachers using standardized procedures provided by the test publisher.

To summarize students' performance across reading, writing, listening, and research domains, the Smarter Balanced ELA test reports a vertically-scaled scale score, which ranges from approximately 2000 to 3000. Test scores are derived based on results of item response theory models, specifically the two-parameter logistic model and the generalized partial credit model, that calibrate items horizontally (within grade) and vertically (between grades) using maximum likelihood estimation methods (SBAC, 2017c). There are no special weights for different domains or different item types. The weighting is achieved via the test blueprint (SBAC, 2017c). Based on predefined grade-specific cut scores, a scale score is converted into an achievement level, which is reported as one of four levels: 1 = Novice, 2 = Developing, 3 = Proficient, and 4 = Advanced. For the purpose of this study, and consistent with prior research (see Furey et al., 2016; Keller-Margulis, Payan et al., 2016; Wilson, 2018), at-risk students were defined as those scoring in the bottom two achievement levels; students scoring in the top two levels were defined as not at risk. Defining risk categories this way was also consistent with accountability requirements for schools, who need to report the percentage of students scoring at level 3 or above (SBAC, 2017b). As shown in Table 1, using this criterion, 40% of the sample in Grades 3 and 4, and 30% in Grade 5, qualified as at risk.

The Smarter Balanced ELA test has been subjected to rigorous content and psychometric validation. Relevant reliability information is reported based on the 2015-16 Technical Report (SBAC, 2017c). Specifically, the marginal reliability of the ELA test, which is defined as one minus the ratio of mean error variance to observed score variance, was .93 for Grade 3, .92 for Grade 4, and .93 for Grade 5. Marginal reliabilities of the ELA test for students with Limited English Proficiency were .88, .87, and .87 for Grades 3, 4, and 5, and marginal reliabilities for students with an IDEA indicator were .90 for all three grades. The overall accuracy of the Smarter Balanced ELA test for classifying students into the achievement level wherein their true score fell was .80 for Grades 3, 4, and 5, indicating good classification accuracy (SBAC, 2017c). Classification accuracy cannot be perfect because the standard error of measurement associated with a student's scale score may result in a 95% confidence interval that spans multiple levels.

Writing screener. The writing screener consisted of the average score of students' responses to six 30min writing prompts scored for holistic quality using the Project Essay Grade (PEG) automated essay scoring (AES) system. Writing prompts addressed three different genres of writing: narrative, informative, and persuasive. These genres were selected because of their prominence within the Common Core state standards and the Smarter Balanced ELA test.

Eighteen prompts (six per genre) were developed by the first author with three certified elementary teachers. Prompt topics were selected based on a high likelihood that students would possess sufficient background knowledge to enable a response. Some of the prompt topics for the narrative genre included: "Write a story about what you would do if you saw an alien," "Write a story about what you would do if you woke up as an animal," and "Write a story about what you would do if you could fly." Some of the prompt topics for the informative genre included: "Think about your favorite place. Teach your reader all about this place," "Think about what it means to

be healthy. Teach your reader different ways to be healthy," and "Think about friendship. Teach your reader the most important things about being a good friend." Some of the prompt topics for the persuasive genre included: "Should kids get to pick their own bed time?" "Should kids get to watch as much TV as they want?" and "Should kids get money for good grades?" The full list of all 18 prompts is provided in Wilson et al. (2019). The writing prompts did not require students to read and integrate additional stimulus material in order to rule out the possibility that low scores on the writing screener were due to weak reading skills. Students who had difficulty reading the prompt were instructed to raise their hand so that the prompt could be read to them.

Across classrooms, we counterbalanced genre order and, within classrooms, we randomly assigned students two prompt topics per genre from a bank of six prompts per genre (18 prompts total) in order to control for order and prompt effects. Prompts were administered using a set of standardized directions that informed students of the purpose of the writing activity, provided students with basic information about the genre (e.g., "A well written persuasive argument tells your opinion. It also includes reasons and evidence to support your opinion."), encouraged students to plan and review their writing, and explained the amount of time they would have to complete the task: 30min. A 30min composing timeframe was selected to ensure that students had sufficient time to plan, draft, and review their essays within a single 45-50min class period after allowing time to distribute materials, read standardized directions, and answer student questions. Though the composing timeframe in the present study was shorter than the 45min timeframe utilized by Wilson (2018), 30min still allowed for eliciting a broader range of cognitive and metacognitive writing processes (see Hayes, 2012) than administration times used for CBM-W screeners (range = 3-10min). Using a 30min timeframe also allowed for extending prior research on the use of AES for screening by experimenting with a slightly shorter composing time.

The PEG scoring system. Student responses were scored for writing quality using the PEG automated scoring and feedback system (Page, 2003) accessible via its web-based formative assessment platform called *PEG Writing.* PEG Writing was rebranded in Fall 2019 as *MI Write* (www.miwrite.net). PEG is proprietary to Measurement Incorporated who has continually updated and improved PEG's scoring capabilities since 2003 (Bunch, Vaughn, & Miel, 2016).

PEG provides automated scores of writing quality using grade-band and genre-specific, but prompt-independent "Six Trait" scoring models. This means that PEG uses different scoring models for different grade bands (Grades 3-4, 5-6, 7-8, 9-10, 11-12) and different scoring models within a grade band to evaluate narrative, informative, and persuasive genres (i.e., 15 scoring models [5 grade bands * 3 genres]). As a prompt-independent scoring model, PEG is designed to assess the Six Traits of writing quality for any prompt topic within a genre, allowing it to be used with teacher- or researcher-designed writing prompts.

To build PEG's scoring models, a training set of data that consists of a corpus of student writing is collected and scored by trained human raters for the following six traits of writing quality on a 1-5 integer scale: elaboration and idea development, organization, style, sentence structure, word choice, and conventions. The next step in the model-building process is feature extraction. PEG measures more than 500 natural language processing (NLP) features that include a combination of features defined a priori based on their relationship with human scores of writing quality (e.g., word frequency, grammatical accuracy, etc.) and features that are identified a posteriori via a machine learning algorithm called support vector regression, which is designed to maximize prediction of the rater scores for each trait (see Smola & Scholkopf, 2004). Though the full NLP feature set used by PEG is proprietary, example NLP features include the following: counts of discourse elements (elaboration and idea development trait); number of transition words,

logical adverbs, cohesion measures (organization trait); usage of pronouns and prepositional phrases (style trait); measures of syntactic complexity and variety (sentence fluency trait); word frequency, counts of technical/sophisticated words and hypernyms (word choice trait); and number and type of spelling and grammar errors (conventions trait).

Once a scoring model is trained to assign scores that are highly consistent with those assigned to the training data, the model is tested with a new set of essays (i.e., a test set). If the PEG scores remain highly consistent with the rater scores for the test set, the model is considered validated and ready for deployment in formative or summative scoring systems.

The PEG holistic quality score. To index a student's overall writing quality, PEG reports a holistic score (henceforth 'PEG Score') that is formed as the sum of the Six Trait scores (range: 6-30). In the present study, all analyses were conducted using the PEG Score because (a) the holistic writing quality score, and not the trait scores, is more aligned with the type of broad, general outcome measures used within universal screening (National Center on Intensive Intervention, n.d.); and (b) psychometrically, the trait scores were highly correlated (range r = .75-.94), loading on a single factor that explained 82% of trait-score variance. This high level of correlation among the trait scores, which most likely replicated correlations among the human scores (see Gansle et al., 2006), suggests that the trait scores would provide similar predictive information; thus, the PEG score was a more suitable measure for the purposes of screening.

PEG is a 100% consistent scoring system with a high level of agreement with human ratings (Shermis, 2014; Shermis et al., 2002). AES-human agreement is commonly reported using the quadratic weighted kappa (QWK) statistic to measure agreement after adjusting for the degree of disagreement: disagreements that are further apart more negatively impact reliability than disagreements that are closer together. QWK values for PEG's scoring models (i.e., 5 grade-bands

* 3 genres * 6 traits) used in its web-based formative assessment system average .834, (*SD* = .026), indicating a highly reliable scoring system.

Word count. In addition to obtaining the PEG Score, we used the word count function in Microsoft Word to calculate the word count (i.e., length) for each student's essay.

Wechsler Individual Achievement Test, 3rd Ed. We used the Sentence Composition subscale of the WIAT-III to provide a standardized, norm-referenced evaluation of students' writing skills. The Sentence Composition subscale is comprised of two individual subtests. The Sentence Combining subtest is comprised of five items that include two or three sentences that are to be combined into a single sentence. The Sentence Building subtest is comprised of seven items with a target word (e.g., *the, or, until*) that must be correctly incorporated into a sentence. Both subtests are scored for semantics/grammar and mechanics. Research assistants trained and then scored these subtests. A single rater scored all the items and a second rater performed a reliability check on a randomly selected 30% sample. Scoring reliability for the Sentence Combining subtest was *r* = .97, with 81% exact + adjacent agreement (i.e., scores that were exact or differing by no more than one point). Scoring reliability for the Sentence Building subtest was *r* = .99, with 86% exact + adjacent agreement. Scoring disagreements were resolved through discussion to generate a final raw score. Raw scores were converted to age-based standard scores. Test-retest reliability of the Sentence Composition subscale for age-based standard scores ranges from .84-.90 across ages 7 to 11, the age span of students in our sample.

Study Design

The present study implemented new analyses to data collected as part of a research project reported in Wilson et al. (2019). The present study addressed different research questions and included additional measures and different data analyses than Wilson et al. (2019). Specifically,

the present study used a retrospective classification design (see STARD, 2015) to evaluate how accurate a multi-prompt writing screener scored using AES in the Fall of 2015 classified students as having failed or passed the Smarter Balanced ELA test administered in Spring 2016.

Study Procedures

In the Fall of 2015, after developing the 18 writing prompts, the first author trained a group of research assistants in a three-hour session to administer study measures, including the WIAT III Sentence Combining and Sentence Building subtests and the researcher-developed writing prompts. Then, data collection proceeded within each school for seven successive days. On Day 1, to collect norm-referenced data on students' writing performance at baseline, research assistants administered the WIAT III subtests to intact classrooms following a set of standardized directions from the WIAT III administration manual. On Days 2-7, research assistants administered the researcher-developed writing prompts to intact classrooms, one prompt per day, following a set of standardized administration directions. Students composed their responses to the six writing prompts on paper rather than on computer because at the time of the research project the schools had insufficient computing resources to enable assessing multiple grade levels concurrently–the district has since transitioned to 1:1 devices. Data collection concluded after the seventh day at which point students had completed the two WIAT III subtests and six 30min writing prompts.

All testing sessions (n = 217) were audio recorded. Thirty percent of the audio recordings were randomly sampled to evaluate fidelity of assessment administration, which was calculated as a percentage of the administration steps completed accurately, including whether or not the research assistant correctly read the set of directions and ensured that students were given exactly 30min to compose their responses to the writing prompts. Fidelity was equal to 98%.

Then, students' handwritten compositions were transcribed verbatim – including errors in grammar, punctuation, and spelling – into Word documents by research assistants. To ensure transcription accuracy, a random sample of approximately 30% of the writing prompts (n = 1,026) were evaluated to calculate the percentage of correctly transcribed elements (i.e., words with original spellings, capitalization, line and paragraph breaks, and punctuation marks). Transcription accuracy was very high: M = 98.51%, SD = 1.64%. Next, the transcribed Word documents were submitted to PEG to for scoring.

Finally, in Spring 2016, school personnel administered the Smarter Balanced ELA assessment and, in Summer 2016, we obtained from the school district students' achievement levels on the Smarter Balanced ELA test. Demographic data and Smarter Balanced data were combined with the PEG Scores in a database that was used for data analysis.

Data Analysis

Research question 1. Our first research question explored how accurate the AES-scored multi-prompt writing screener was for making classification decisions. To answer this question, we first calculated students' average PEG Score (range = 6-30) across the six prompts. This meant averaging PEG scores from prompts administered in three genres with two different, randomly assigned topics per genre.

Overall classification accuracy. The overall classification accuracy of the average PEG score was estimated via the area under the receiver operating characteristic (ROC) curve. A ROC curve is a plot of the sensitivity (i.e., true positive rate) and false positive rate for each unit comprising the scale of a predictor variable. In the current context, the area under the curve (AUC) summarizes the probability that a randomly selected at-risk student would perform worse on the writing screener than a randomly selected not at-risk student (Zweig & Campbell, 1993). AUC

values range from .50 – 1.00 and can be interpreted as follows: .50 = chance; .50 – .70 = poor accuracy; .70 – .80 = acceptable accuracy; .80 – .90 = good accuracy; > .90 = excellent accuracy (Hosmer, Lemeshow, & Sturdivant, 2013). We adopted an AUC of .75 as a minimal level of accuracy to show the promise of a screening model (Cummings & Smolkowski, 2015), and .85 as a desirable level of accuracy for implementation in schools (Christ & Nelson, 2014).

Threshold-dependent measures of classification accuracy. Having estimated the ROC curve, it was possible to select a cut score within the range of the possible PEG Scores to optimize different classification decisions. Selecting a cut score involves balancing sensitivity (i.e., true positive rate) and specificity (i.e., true negative rate). Sensitivity and specificity describe how well a cut score classifies students overall, at the group level, such as an entire grade-level within a school or district (Van Norman, Klingbeil, & Nelson, 2017). Sensitivity is calculated as: true positives/(true positives + false negatives). Specificity is calculated as: true negatives/(true negatives + false positives). Consistent with recent recommendations regarding advances to screening research (VanDerHeyden & Burns, 2018), we selected multiple cut scores using different criteria and provided sensitivity and specificity values for each cut score.

First, we explored a *d*-based cut score (Swets & Picket, 1982), which maximizes sensitivity and specificity. *d* represents the point on the ROC curve that minimizes the distance between the curve and the upper left-hand corner of the ROC space, a point that represents perfect classification. The *d* statistic is calculated as $\sqrt{(1\text{-sensitivity})^2 + (1\text{-specificity})^2}$. When evaluating the *d*-based cut score, we considered sensitivity values of ≥ .80 as "acceptable," and values in the range .70-.80 to be "borderline acceptable" (Kilgus, Chafouleas, & Riley-Tillman, 2013; Silberglitt & Hintze, 2005). Second, we explored two sensitivity-based cut scores. In most school-based screening research, sensitivity is prioritized over specificity because it is assumed that failing

to identify students who are at risk (i.e., a false negative) is more costly than a false positive. We explored cut scores where minimal sensitivity was set to 90% (see Jenkins et al., 2007) and 80% (see Smolkowski & Cummings, 2015). We examined concomitant specificity levels for all three cut scores (i.e., the *d*-based cut score, and the two sensitivity-based cut scores). High levels of specificity are important so that school resources are not overburdened because of false positive errors. We considered specificity values of .80 as "desirable," .70 as "acceptable," and values between .60-.70 as "borderline acceptable" (Kilgus et al., 2013; Silberglitt & Hintze, 2005).

We calculated posttest probabilities to complement sensitivity and specificity. Prior research on writing screeners has not rigorously evaluated screening models with respect to posttest probabilities. Thus, in keeping with current advances in screening research (Pendergast, Youngstrom, Ruan-Lu, & Beysolow, 2018; VanDerHeyden, 2013; Van Norman et al., 2017), we calculated positive and negative posttest probability values for each cut score.

Positive posttest probability (PP+) values represent the probability that a student identified as at risk would actually go on to fail the outcome assessment, which is akin to the probability of being a true positive. Negative posttest probability (PP-) values represent the probability that a student identified as not at risk goes on to fail the outcome assessment, which is akin to the probability of being a false negative. The PP+ is calculated as

$$\frac{\left(\frac{pretest\ probability}{1-pretest\ probability}\right) * LR +}{\left(\frac{pretest\ probability}{1-pretest\ probability} * LR +\right) + 1}$$

The LR+ (the positive likelihood ratio) is calculated as sensitivity/(1-specificity). The PP- is calculated similarly except the LR+ is substituted for the LR- (the negative likelihood ratio), which is calculated as (1-sensitivity)/specificity. In the current study, we used the base rate (within

a grade) as the pretest probability, which meant that the pretest probability was consistent across students, a practice consistent with prior research (e.g., Van Norman et al., 2017).

VanDerHeyden (2013) suggested the following interpretations of posttest probability values: <.10 = safe to assume that a student does not require intervention; .10-.50 = administer additional assessments to determine risk status; >.50 = provide intervention. Thus, ideally, a cut score used for a screener should yield a PP+ value >.50 and a concomitant PP- value of ≤ .10, which means that a student identified as at risk has more than a 50% chance of going on to fail the outcome measure and a student identified as not at risk using that cut score has less than or equal to a 10% chance of being a false negative.

Finally, we calculated the difference in posttest and pretest probabilities. The difference between these probabilities signifies the "added value," so to speak, of the screener over relying on our knowledge of preexisting risk, which in this case was the base rate (Van Norman et al., 2017). In other words, if a base rate is 40%, then we assume that each student has a 40% chance of being at risk. If a screener yields a PP+ of .55 and a PP- of .15, this means that using the screener increases the probability of correctly identifying a truly at-risk student by 15% and reduces the probability of making a false negative decision by 25%. Accordingly, we calculated the difference in posttest and pretest probabilities as: PP+ – base rate and PP- – base rate.

Research question 2. Research question 2 considered whether a similar level of overall classification accuracy could be achieved between a PEG Score derived from the average of six prompts and a PEG Score derived from the average of fewer prompts, or even a single prompt. To answer this research question we calculated the average PEG Score of the first five prompts (a five prompt screening method), the first four prompts (a four prompt screening method), and so on until only a single PEG Score remained. This single score corresponds to the score students

received for the first writing prompt they were administered. This process yielded five new scores for each student, which were then used as predictors in five new ROC curve analyses. We compared AUC values for these screening models to the AUC value of the six-prompt screener. Then, we tested whether observed differences in AUC values were statistically significantly different using the Hanley and McNeil (1983) method for comparing AUC values drawn from the same sample. The method calculates a *z*-score for the difference between two dependent AUC values using the following equation: $z = (\mathrm{AUC}_1 - \mathrm{AUC}_2)/\sqrt{(SE_1^2 + SE_2^2) - (2(r) * SE_1 * SE_2)}$, where *r* is the estimated correlation between the two AUC values. Based on the results of those comparisons, if a more efficient screener yielded a non-statistically significantly different AUC value as that of the six-prompt screener, we calculated additional classification accuracy measures using the *d*-based and sensitivity-based cut scores.

Research question 3. Research question 3 compared whether classification accuracy was superior when the screener was scored using AES or using word count. Accordingly, we calculated AUC values for the writing screener scored for word count and used Hanley-McNeil comparisons to compare those AUC values to ones obtained when using AES.

Multiple imputation. Multiple imputation using Markov Chain Monte Carlo algorithms was conducted to handle missing data. The percent of missing data for the six writing screeners ranged from 4.6% – 6.9% and was 5.8% for the state ELA test. The multiple imputation model included the following variables as predictors: grade, classroom, school, gender, race, special education status, ELL status, and chronological age. The PEG score, the Smarter Balanced ELA scale score, and Smarter Balanced achievement level were entered in the model as both predictors and the foci of imputation. Five datasets were generated in addition to the original data, and these five datasets were used in all analyses. Pooled results were calculated by hand using the following

formulas from Rubin (1987). The combined point estimate for the AUC value ($\bar{Q}$) was calculated as follows: $\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$, where *m* is the number of imputed data sets and $\hat{Q}_i$ is the AUC estimate for the *i*th imputed data set. The corresponding combined standard error estimate was calculated as: $SE = \sqrt{\bar{W} + \left(1 + \frac{1}{m}\right) * B}$, where $\bar{W}$ is the within-imputation variance and B is the between-imputation variance. $\bar{W}$ is calculated as $\frac{1}{m}\sum_{i=1}^{m}\widehat{w_i}$, where $\widehat{W_i}$ is the nonparametric standard error estimate for the *i*th imputed data set. *B* is calculated as $\frac{1}{m-1}\sum_{i=1}^{m}(\widehat{Q_i} - \bar{Q})^2$

Statistical software. SPSS V.25 was used to conduct multiple imputation, estimate descriptive statistics and correlations, and estimate the ROC curves. Microsoft Excel was used to calculate pooled results, Hanley-McNeil comparisons, posttest probabilities, and differences between posttest and pretest probabilities. All analyses were conducted separately by grade level.

Results

Correlations and Descriptive Statistics

Tables 2, 3, and 4 present point-biserial correlations for the PEG Score with Risk Status (coded 1 = at risk and 0 = not at risk) as well as descriptive statistics for Grades 3, 4, and 5, respectively. The first row of each table presents the PEG Score resulting from the average of six prompts. Subsequent rows of each table present the PEG Score resulting from the average of one fewer writing prompt concluding with students' score on the first prompt they were administered.

Key findings include the following. First, for all grades, there were statistically significant point-biserial correlations between the PEG score for each screening composite and Risk Status. All correlations were in the expected direction: higher PEG Scores were associated with a decreased likelihood of being at risk. Second, the correlations with Risk Status among the PEG Scores for the different screening composites were generally similar for Grade 3 (range *r* = -.44 –

- .47), fluctuated slightly for Grade 4 (range r = -.42 – -.49), and fluctuated the most for Grade 5 (range r = -.33 – -.45).

Third, there appears to have been a floor effect when examining the means and observed range for the PEG Scores in all grades. Fourth, though not reported in the tables for brevity, comparisons of PEG scores by gender, special education status, and English language learner (ELLs) status, indicated that males, students with disabilities, and ELLs all scored statistically significantly lower (p < .05) than their counterparts on the different PEG screeners. These differences in PEG scores across subgroups are consistent with well-established performance differences in the writing area (Deane, 2018). Finally, z-scores for skewness (skewness/*S.E.*) and kurtosis (kurtosis/*S.E.*) show that the distribution of the PEG score was within normal bounds, as indicated by z-scores ≤ |3.29|, corresponding to an alpha of 0.001 (Tabachnick & Fidell, 2013).

Research Question 1: Classification Accuracy of the AES Screener

Overall classification accuracy. The area under the ROC curve (AUC) and its associated 95% confidence interval (CI) for the average PEG Score of six writing prompts for Grade 3 was .78 (SE = .034; 95% CI [.72, .85]). For Grade 4 the AUC was .79 (SE = .036; 95% CI [.72, .86]). For Grade 5 the AUC was .81 (SE = .035; 95% CI [.74, .87]). In each grade, the AUC for the average PEG Score of six writing prompts exceeded the .75 criterion for minimal acceptable accuracy to show the promise of a screening model (Cummings & Smolkowski, 2015). No AUC met the .85 criterion for desirable accuracy (Christ & Nelson, 2014).

Sensitivity and specificity. Table 5 summarizes sensitivity and specificity values associated with different cut scores for the average PEG Score of the six-prompt screener for Grades 3, 4, and 5. Three cut scores were used: a d-based cut score that optimizes sensitivity and specificity, and two cut scores that fix sensitivity at either 90% or 80%.

For Grade 3, none of the cut scores yielded simultaneously acceptable sensitivity and specificity values. For instance, the *d*-based cut score yielded sensitivity of .66 and specificity of .80, indicating unacceptable sensitivity but desirable specificity. When fixing sensitivity to 90% and 80%, the associated specificity levels were unacceptable. Results were similar for Grade 4. The *d*-based cut score yielded unacceptable sensitivity (.64), but desirable specificity (.83). When fixing sensitivity to 90% and 80%, specificity was unacceptable (< .60). Results were improved for Grade 5. The *d*-based cut score yielded acceptable sensitivity and specificity values: sensitivity was .80 and specificity was .73. When fixing sensitivity to 90% and 80%, the associated specificity levels were unacceptable.

Posttest probabilities. For Grade 3 using a *d*-based cut score, a student identified as at risk would have a 69% probability of truly being at risk and a student identified as not at risk would have a 22% probability of being a false negative. The 90% and 80% sensitivity-based cut scores resulted in a student having a 54% probability of being at risk (PP+), which exceeds the 50% threshold suggested by VanDerHeyden (2013). These sensitivity-based cut scores would result in greater confidence in ruling out false negative errors (PP-): the cut scores resulted in either a 12% or 17% probability of a student being a false negative. Examining the gains in accuracy over relying on pretest probabilities alone (i.e., the base rate), it is clear that the *d*-based cut score provides the greatest gain in PP+ but the lowest gain in PP-. Specifically, using the *d*-based cut score for the average PEG Score of six writing prompts would increase the probability of correctly identifying a truly at risk student by 29% and would reduce the probability of making a false negative decision by 18%. Comparable results were found for Grade 4.

For Grade 5 students, PP+ values for the *d*-based cut score, the 90% cut score, and 80% cut score were all above the .50 criterion; PP- values for these cut scores were .11, .13, and .15,

respectively. Examining the gains in PP+ over pretest probabilities, the three cut scores yielded similar increases over the base rate in the probability of correctly identifying a truly at risk student, 16%, 13%, and 17%, respectively. Using these cut scores would also result in a sizeable reduction in the probability of making a false negative decision, 29%, 27%, and 25%, respectively.

Research Question 2: Examining Efficiency

For Grade 3, Hanley-McNeil comparisons indicated that the AUC value of the average PEG Score of six prompts was not statistically significantly different than the average PEG Score of five prompts (z = 0.24, p = .813), four prompts (z = 0.30, p = .766), three prompts (z = 0.93, p = .351), two prompts (z = 0.81, p = .417), or students' PEG Score on their first prompt (z = 0.94, p = .347). The same was true in Grade 4. Hanley-McNeil comparisons indicated that the AUC value of the average PEG Score of six prompts was not statistically significantly different than the average PEG Score of five prompts (z = -0.23, p = .819), four prompts (z = 0.23, p = .820), three prompts (z = -0.34, p = .731), two prompts (z = 1.40, p = .162), or students' PEG Score on their first prompt (z = 1.23, p = .219). Thus, the most efficient method of using PEG to accurately classify third and fourth graders as having failed or passed the Smarter Balanced ELA test was to use a screener consisting of a single writing prompt in a randomly assigned genre with a randomly assigned topic. The AUC value of the single prompt screener in Grade 3 was .76 (*SE* = .037; 95% CI [.69, .83]) and in Grade 4 was .76 (*SE* = .038; 95% CI [.68, .83]).

In Grade 5, the AUC value of the average PEG Score of six prompts was not statistically significantly different than the average PEG Score of five prompts (z = 0.29, p = .770), four prompts (z = -0.18, p = .860), or three prompts (z = 0.74, p = .459). However, the AUC associated with the average PEG Score of six writing prompts was statistically significantly greater than the AUC associated with the average of two prompts (z = 2.75, p = .006) and the score for a single

writing prompt (z = 2.67, p = .008). Thus, the most efficient method of using PEG to accurately classify fifth graders as having failed or passed the Smarter Balanced ELA test was to use a screener consisting of three writing prompts in randomly assigned genres with randomly assigned topics. The AUC value of the average PEG Score for the three-prompt screener in Grade 5 was .79 (*SE* = .036; 95% CI [.72, .86]).

Table 6 presents additional classification accuracy statistics for these "most efficient" PEG screening methods. For each grade, only the *d*-based cut score and 90% sensitivity-based cut score are presented; no cut scores successfully achieved 80% sensitivity. For all grades, only the *d*-based cut score yielded minimally acceptable classification accuracy, as indicated by sensitivity and specificity values above .70. For the 90% sensitivity-based cut score, specificity was unacceptable.

With respect to posttest probability values, for Grades 3 and 4, the *d*-based cut score would generate PP+ values above .50, which is optimal, but PP- values were .19, higher than the desired .10 criterion suggested by VanDerHeyden (2013). The single-prompt screener scored using PEG would increase the accuracy with which individual students were classified as at risk by 26% over their pretest probability. It would also decrease the probability of making a false negative decision by 21%. Both are favorable outcomes for screening models.

Finally, in Grade 5, the *d*-based cut score would generate a PP+ value above .50, which is optimal, and a PP- value of .12, which approaches the desired .10 criterion suggested by VanDerHeyden (2013). Screening students using the average PEG Score of three writing prompts would increase the probability of making true positive decisions by 14% and decrease the probability of making false negative errors by 28%.

In sum, the classification accuracy measures reported in Table 6 indicate that acceptable, and comparable, classification accuracy can be achieved when using these "most efficient" PEG screening models versus the full six-prompt screening model (see Table 5).

Post-hoc analysis. Results of RQ2 indicated that differing numbers of writing screeners across grades were needed to achieve a commensurate level of classification accuracy to the six-prompt screener: a single prompt was all that was needed in Grades 3 and 4, but three prompts were needed in Grade 5. Why would results differ by grade?

A plausible explanation is that third- and fourth-grade students demonstrated a more stable rank order across prompts than fifth-grade students. Greater stability in rank ordering across prompts, especially at or near the cut score, would allow for the use of screening with fewer scores. Alternatively, greater variability in rank order would necessitate the use of more scores to achieve a more reliable measure, which in the current study was attained in Grade 5 by averaging scores from three writing prompts.

As a post hoc test of this hypothesis, we examined whether the correlation between the PEG Score derived from a single prompt (i.e., the first prompt assigned) and the average of all six writing prompts was greater in Grades 3 and 4 than in Grade 5. Specifically, we estimated this correlation in all three grade levels and then compared the correlations using Fisher's *r*-to-*z* transformation. We hypothesized that the correlation would not be statistically significant different between Grades 3 and 4 and used a two-tailed test of this hypothesis. We also hypothesized that the correlations in Grades 3 and 4 would be greater than the correlation in Grade 5, and thus used a one-tailed test of this hypothesis.

The correlation between the first prompt and the average of all six prompts in Grade 3 (n = 185), Grade 4 (n = 167), and Grade 5 (n = 187) was .84, .80, and .72, respectively. Results of

Fisher's *r*-to-*z* transformation test indicated that the correlations between the first prompt and the average of all six prompts for Grades 3 and 4 were not statistically significantly different ($z = 1.14$, $p = 0.254$), but the correlation was greater in Grade 3 than Grade 5 ($z = 3.00$, $p = 0.001$) and greater in Grade 4 than Grade 5 ($z = 1.78$, $p = 0.038$). Results lend partial support for the hypothesis that more stable rank orderings in Grades 3 and 4, and less stable rank orderings in Grade 5, were related to the ability to use a single prompt for screening in Grades 3 and 4—a single prompt was more representative of the average of six prompts in these grades—and the need to use a multi-prompt screener in Grade 5 where a single prompt was less representative of a student's average performance.

Research Question 3: Comparing AES with Word Count

Consistent with prior research (Deane, 2013; Perelman, 2014), we found high correlations between AES and word count for the six-prompt screener for Grade 3 ($r = .85$), Grade 4 ($r = .79$), and Grade 5 ($r = .90$). High correlations were also evident between AES and word count for the "most efficient" screener; that is, the single-prompt screener in Grades 3 and 4 ($r = .78$ and .76, respectively) and the three-prompt screener in Grade 5 ($r = .89$). However, as indicated in Table 7, AUCs for the screeners scored using word count were statistically significantly worse than those obtained using the PEG Score and tended to fall in the "poor" range. Despite high correlations with word count, classification accuracy was superior when screening students with AES.

Discussion

Given the challenges facing writing screening, and the limitations of existing approaches to writing screening (e.g., CBM-W), it is important to examine alternative writing screening procedures that are efficient to administer and score and hold promise for accurately classifying students into risk categories. The present study explored a proof of concept of writing screening using automated essay scoring (AES) software, an increasingly prevalent form of educational

technology in U.S. schools (Stevenson & Phakiti, 2014; Wilson & Czik, 2016; Wilson & Roscoe, 2020).

Based on prior research that found that generalizable low-stakes decisions could be made about struggling writers using the average PEG Score of six writing prompts (Wilson et al., 2019), our first research question explored the classification accuracy of a six-prompt writing screener. Results indicated that overall classification accuracy of the six-prompt screener was acceptable, though not desirable, and cut scores could be selected that yielded acceptable classification decisions, as measured by sensitivity, specificity, and positive and negative posttest probabilities. Our second research question tested whether a more efficient form of that screener, one consisting of fewer than six writing prompts, could be used without sacrificing classification accuracy. Results showed that similar and acceptable levels of classification accuracy could be achieved with a single writing prompt in Grades 3 and 4 and with the average score of three writing prompts in Grade 5. In all cases, comparisons of the PP+ and PP- values to the base rate indicated that screening students for risk of failing the Smarter Balanced ELA test using writing prompts scored via PEG improved the accuracy screening decisions. Finally, our third research question tested whether screening using AES yielded greater classification accuracy than screening using word count, a more common and less expensive measure. Despite high correlations between AES and word count, classification accuracy was consistently and significantly higher when using AES.

Findings add to a small but growing body of research that shows the promise of AES screening methods as an alternative to human-scored methods like CBM-W (Wilson, 2018; see also Mercer, Keller-Margulis, Faith, Reid, & Ochs, 2018). Results showed that the PEG Score for a single 30-min writing prompt in Grades 3 and 4, and the average PEG Score across three 30-min writing prompts in Grade 5, yielded acceptable classification accuracy. In the present study, the

AUC for the single prompt screeners in Grades 3 and 4 was .76, and for the three-prompt screener in Grade 5 the AUC was .79. AUCs were similar to those of Wilson (2018), who reported an AUC value of .74 for Fall Grade 3 and .79 for Fall Grade 4. AUCs also were in the same range as prior studies of CBM-W in the upper-elementary grades: Furey et al. (2016) reported AUCs in the range of .74-.80 for Grade 4 students, and Keller-Margulis, Payan et al. (2016) reported AUCs in the range of .61-.78 for CBM-W measures administered in the fall of Grade 4.

It is important to note that the length of CBM-W administration is much shorter (3-10min) than the administration times used for the writing prompts in the present study (30min). On the surface, the use of a 30min administration time may appear to impinge on the efficiency of AES for screening. However, this is likely not the case when factoring in the time associated with scoring CBM-W. Although Gansle et al. (2004) estimated the average time to score a 3-min CBM-W probe for correct word sequences at just over a minute, others have estimated the average time to be in the range of 4-5min per composition (Espin et al., 1999). Given a classroom of 25 students, the total time to administer and score a CBM-W prompt would be approximately 4min to administer (1min to plan, 3min to write) and 25-100min to score, totaling somewhere in the range of 29-104min. However, given problems with the reliability of CBM-W, it is recommended that at least three 3min prompts or two 5-7min writing prompts be administered to make reliable relative decisions about students (Keller-Margulis, Mercer, & Thomas, 2016), which places the total time in the range of 87-312min. Thus, adopting a 30min writing prompt scored using AES is, at worst, just as efficient and, more likely, far more efficient than CBM-W. Indeed, even if students were to complete three 30min writing prompts scored using AES, the total administration and scoring time for a classroom, or an entire district for that matter, would be only 90min.

Further, there is a pedagogical benefit of longer administration times. Requiring students to write for 30min, and not 3min, better aligns with recommendations for strengthening Tier 1 instruction (Graham, Bollinger, et al., 2012; Graham, McKeown, et al., 2012). Due in part to the time costs of scoring writing, teachers rarely assign writing that requires an extended response (i.e., greater than a single paragraph; Graham, McKeown, et al., 2012). Yet, college and career readiness standards emphasize the need to build writing stamina and develop metacognitive control over writing processes (National Governors Board and Council of Chief State School Officers, 2010). Thus, time spent writing adds to, rather than detracts from, instruction.

Screening Efficiency Differs by Grade Level

The most efficient writing screener differed by grade level: a single prompt screener was sufficient for Grades 3 and 4, but a three-prompt screener was necessary for Grade 5. Our post hoc analysis indicated that the correlation between the score on the first prompt and the average score of all six prompts was non-statistically significantly different in Grades 3 and 4, but statistically significantly greater in Grades 3 and 4 than in Grade 5. Stability and variability of rank orderings may, in part, help to explain why screening efficiency was found to differ by grade level: A single prompt is more representative of a student's average performance across six prompts in Grades 3 and 4 than it is in Grade 5.

While the present study is not positioned to determine reasons why fifth graders displayed greater variability in rank order across prompts, research suggests that genre knowledge and topic knowledge increase across grade-levels (Olinghouse & Graham, 2009) and that writing interest and motivation decrease (Pajares, 2003). Therefore, it is plausible that at-risk fifth graders may be more susceptible to prompt effects because of complex interactions between the writing prompt and students' genre knowledge, topic knowledge, and prompt-specific motivation. Indeed, a prior

generalizability study of PEG scores (Wilson et al., 2019) revealed that residual variance associated with interactions between students, prompts, and genre was higher in Grade 5 than in Grades 3 and 4. As such, greater variability in rank ordering and performance may be related to our finding that screening fifth graders required a greater number of writing prompts to make accurate screening decisions. Additional research is needed to further explore this finding.

AES and Word Count

We found that the PEG Score and word count were highly correlated, but classification accuracy was superior when writing prompts were scored using PEG. It is well-documented that AES and word count are highly correlated (see Deane, 2013), which critics use as evidence of lack of construct validity (Perelman, 2014). Yet human-scored measures of writing quality are also highly correlated with word count (Perelman, 2012), particularly in the elementary grades because students who generate an insufficient amount of text tend to score low on virtually all dimensions of writing quality (e.g., development of ideas, organization, word choice, sentence fluency, etc.). Thus, high correlations with text length, especially in the elementary grades, are not in and of themselves cause for dismissing AES. In fact, AES may exhibit high correlations with text length without directly measuring word count; many text features included in AES algorithms, such as the number of discourse elements or number of complex sentences, are also correlated with text length (Deane, 2013).

Nevertheless, the use of AES comes with a cost. If word count were to classify students as well as AES, the costs of AES would not be justified for universal screening. Yet, findings of the present study indicate that word count cannot replace AES as a screener. These findings are consistent with those of a prior study of AES for predicting middle school students' state test performance (Wilson et al., 2016), and are consistent with broader findings from the CBM-W

literature that word count is among the weakest screening measures (see Furey et al., 2016; Keller-Margulis, Payan et al., 2016). Findings of the present study do not resolve issues pertaining to the construct validity of AES, but they do provide validity evidence supporting the use of AES for screening students for risk of failing a widely used Common-Core-aligned ELA proficiency test.

Limitations and Directions for Future Research

Study findings must be interpreted in light of the following limitations, each of which suggests an important direction for future research. First, though the PEG system is designed to have students type their essays, limited computer resources at the schools combined with the need to collect data from entire grade levels concurrently, meant that students handwrote their compositions. The compositions were then transcribed verbatim and input into PEG for scoring. In spite of high levels of transcription accuracy, study procedures differed from those that would be recommended if universal screening with AES were to be put into practice. To truly capitalize on the efficiency of AES, it is essential that students type their compositions directly into the AES software. It is simply too inefficient to transcribe compositions for AES scoring. Thus, the current study does not illustrate the intended implementation of AES for screening but offers a proof of concept for doing so. Future research on writing screening with AES should require students to keyboard their responses. This should be feasible considering that all but one state in the U.S. has adopted computer-based assessments (https://nces.ed.gov/programs/statereform/tab2_22.asp) and districts, even those without 1:1 devices, are accustomed to sharing computing resources to ensure that all students are assessed. In light of differences in handwriting and keyboarding performance among elementary students (see White, Kim, Chen, & Liu, 2015), future research should compare the classification accuracy of AES screeners administered via computer versus AES screeners applied to handwritten prompts.

Second, we examined whether performance on a writing screener could be used to accurately classify students at risk of failing the Smarter Balanced English language arts (ELA) test. Our decision to use this outcome measure was influenced by the test's rigorous validation, its widespread use with millions of students across many U.S. states, and the fact that the ELA achievement level has been validated for determining whether or not students met grade-level expectations for writing and literacy skills. Nevertheless, there are two limitations with respect to the use of the Smarter Balanced ELA test as our outcome measure. First, the fact that the Smarter Balanced ELA achievement level represents performance relative to a broader literacy construct, and not just writing, may be considered a study limitation. An area of future research would be to explore combining PEG-scored writing screeners with other measures that predict risk of literacy failure (e.g., reading CBM). One way to do this would be to calculate posttest probability values using pretest probabilities generated from the reading screener. Indeed, in the present study we calculated posttest probabilities by utilizing a pretest probability value, the base rate, that did not vary across individuals within a grade. This precluded the use of our posttest probability values from offering truly individualized information. Calculating posttest probabilities by combining information from reading and writing screeners may offer a potentially more accurate estimate of individual risk of failing the Smarter Balanced ELA test.

Furthermore, similar to most state writing assessments and norm-referenced writing assessments, the Smarter Balanced ELA test assesses student writing using a single extended performance task (i.e., single writing prompt). Yet, results of prior generalizability studies show that obtaining a highly reliable estimate of students' writing ability requires the use of multiple writing prompts and multiple raters; a single writing prompt is insufficient (Graham et al., 2016; Kim et al., 2017; Wilson et al., 2019). The present study examined ways to improve reliability of

screening assessments by obtaining a much larger sample of student writing than in prior studies. However, issues surrounding reliability of measurement apply not only to writing screeners, they also apply to higher-stakes assessments. Thus, future research should consider ways of improving the reliability of state and norm-referenced writing assessments; such efforts have the potential to not only yield more accurate estimates of students' writing proficiency but to also provide more reliable assessments as outcome measures for validating writing screeners.

Third, students were required to incorporate supplemental reading material in the Smarter Balanced ELA writing prompt, but our writing prompts could be answered using students' background knowledge, similar to the types of prompts employed in CBM-W. Differences in prompt types may also explain study findings. Thus, future research should examine classification accuracy of screening models when administering writing prompts that require the use of reading materials. The use of such prompts would incorporate reading skills into the writing evaluation to a greater extent, and potentially improve classification accuracy for predicting state ELA test performance.

Fourth, the 30-min composing timeframe may have exacerbated the observed floor effect in students' PEG scores (see Tables 2-4). Extending the composing time to 45min for students in these grades, as was done in Wilson (2018) who observed less severe floor effects, may be beneficial psychometrically and diagnostically. However, potential gains in accuracy must be weighed against concomitant decreases in the efficiency of prompt administration.

A final limitation pertains to the sample. Consistent with universal screening procedures in which all students within a grade-level are administered the same screening measure, and consistent with prior screening research conducted with diverse samples (Keller-Margulis, Ochs, Reid, Faith, & Schanding, 2019; Kent, Wanzek, & Yun, 2019; Stevenson, 2017; Wilson, 2018),

we elected to retain ELLs and students with disabilities in our grade-level samples. In the present study there was insufficient power to accurately estimate classification accuracy models with ELLs or students with disabilities alone. However, future research should examine whether PEG is equally accurate at classifying ELLs and SWDs using separate, fully-powered samples.

Conclusion: Implications for Researchers and Educators

Findings add to the growing body of research showing the promise of automated scoring for writing screening. AES-scored writing prompt screeners yield acceptable classification accuracy, are efficient, more accurate than screeners scored for word count, and elicit a broader range of cognitive and metacognitive writing processes via the use of longer composing times. Findings also reveal the importance of obtaining a broader sample of students' writing performance in procedures used for universal screening. In particular, as students approach the middle grades, it may be necessary to screen students based on their performance on multiple prompts that assess multiple genres and topics. In sum, the use of automated scoring shows promise for addressing persistent challenges and advancing the field of writing screening.

References

Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in Grades 1 to 7. *Journal of Educational Psychology, 102,* 281-298. DOI: 10.1037/a0019318

Ahmed, Y., Wagner, R. K., & Lopez, D. (2014). Developmental relations between reading and writing at the word, sentence, and text levels: A latent change score analysis. *Journal of Educational Psychology, 106,* 419-434. doi:10.1037/a0035692

Allen, A. A., Poch, A. L., & Lembke, E. S. (2018). An exploration of alternative scoring methods using curriculum-based measurement in early writing. *Learning Disability Quarterly, 41,* 85-99. DOI: 10.1177/0731948717725490

Behizadeh, N., & Pang, M. E. (2016). Awaiting a new wave: The status of state writing assessment in the United States. *Assessing Writing, 29,* 25–41. http://dx.doi.org/10.1016/j.asw.2016.05.003

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.

Berninger, V. W. (2000). Development of language by hand and its connections with language by ear, mouth, and eye. *Topics in Language Disorders, 20*(4), 65-84.

Berninger, V. W., & Abbott, R. D. (2010). Listening comprehension, oral expression, reading comprehension, and written expression: Related yet unique language systems in Grades 1, 3, 5, and 7. *Journal of Educational Psychology, 102,* 635-651. DOI: 10.1037/a0019319

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32,* 83-100. DOI: 10.1177/0265532214542994

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Brindle, M., Graham, S., Harris, K.R., & Hebert, M. (2016). Third and fourth grade teacher's classroom practices in writing: A national survey. *Reading & Writing, 29*, 929–954. DOI 10.1007/s11145-015-9604-x

Bunch, M. B., Vaughn, D., & Miel, S. (2016). Automated scoring in assessment systems. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Technology tools for real-world skill development* (pp. 611–626). Hershey, PA: IGI Global.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on early identification of reading disabilities. *Journal of Learning Disabilities, 42*, 163–176. https://doi.org/10.1177/0022219408326219

Christ, T. J., & Nelson, P. M. (2014). Universal screening in education settings: Evidence-based decision making for schools. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in education settings: Evidence-based decision making for schools* (pp. 79-110). American Psychological Association.

Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). *An investigation of the impact of the 6+1 Trait Writing model on grade 5 student writing achievement* (NCEE 2012–4010). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Cummings, K. D., & Smolkowski, K. (2015). Selecting students at risk of academic difficulties. *Assessment for Effective Intervention, 41*, 55–61. https://doi.org/10.1177/1534508415590396

Davidson, M., & Berninger, V. (2016). Informative, compare and contrast, and persuasive essay composing of fifth and seventh graders: Not all essay writing is the same. *Journal of Psychoeducational Assessment, 34,* 311-321. https://doi.org/10.1177/0734282915604977

Deane. P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18,* 7-24. https://doi.org/10.1016/j.asw.2012.10.002

Deane, P. (2018). *Achievement gaps in writing: Identifying problems, enabling solutions.* Unpublished manuscript, Princeton, NK: Educational Testing Service.

Deno, S. L., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children, 48,* 368–371. https://doi.org/10.1177/001440298204800417

Espin, C. A., Scierka, B. J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading & Writing Quarterly, 15,* 5-27. https://doi.org/10.1080/105735699278279

Espin, C. A., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *The Journal of Special Education, 34,* 140–153. https://doi.org/10.1177/002246690003400303

Espin, C., 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. https://doi.org/10.1177/001440290807400203

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist, 35,* 39-50. https://doi.org/10.1207/S15326985EP3501_5

Furey, W. M., Marcotte, A. M., Hintze, J. M., & Shackett, C. M. (2016). Concurrent validity and classification accuracy of curriculum-based measurement for written expression. *School Psychology Quarterly, 31*, 369–382. https://doi.org/10.1037/spq0000138

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Slider, N. J., Hoffpauir, L. D., & Whitmarsh, E. L. (2004). An examination of the criterion validity and sensitivity to brief intervention of alternate curriculum-based measures of writing skill. *Psychology in the Schools, 41*, 291-299. https://doi.org/10.1002/pits.10166

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression of elementary school students. *School Psychology Review, 35*, 435–450.

Gilbert, J., & Graham, S. (2010). Teaching writing to elementary students in Grades 4-6: A national survey. *Elementary School Journal, 110*, 494-518. https://www.jstor.org/stable/10.1086/651193

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. New York, NY: College Entrance Examination Board.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*, 257-288. https://doi.org/10.1207/S1532799XSSR0503_4

Graham, S., Bollinger, A., Booth Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., & Olinghouse, N. (2012). *Teaching elementary school students to be effective writers: A practice guide* (NCEE 2012- 4058). Washington, DC: National Center for Education

Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/ wwc/publications_reviews.aspx#pubsearch.

Graham, S., & Hebert, M. A. (2010). *Writing to read: Evidence for how writing can improve reading. A Carnegie Corporation Time to Act Report*. Washington, DC: Alliance for Excellent Education.

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,* 72–82. https://doi.org/10.1177/0731948714555019

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*, 879-896. https://doi.org/10.1037/a0029185

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839–843. https://doi.org/10.1148/radiology.148.3.6878708

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*, 369–388. https://doi.org/10.1177/0741088312451260

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression. 3rd Ed.* Hoboken, NJ: John Wiley & Sons, Inc.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263. https://doi.org/10.3102/00346543060002237

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*, 582–600.

Jordan, N. C., Gluting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review, 39*, 181–195.

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*, 383–392. https://doi.org/10.1037/spq0000126

Keller-Margulis, M. A., Ochs, S., Reid, E. K., Faith, E. L., & Schanding, G. T. (2019). Validity and diagnostic accuracy of early written expression screeners in kindergarten. *Journal of Psychoeducational Assessment, 37*, 539-552. https://doi.org/10.1177/0734282918769978

Keller-Margulis, M., Payan, A., Jaspers, K. E., & Brewton, C. (2016). Validity and diagnostic accuracy of written expression curriculum-based measurement for students with diverse language backgrounds. *Reading and Writing Quarterly*, *32*, 174–198. https://doi.org/10.1080/10573569.2014.964352

Kent, S. C., Wanzek, J., & Yun, J. (2019). Screening in the upper elementary grades: Identifying fourth-grade students at-risk for failing the state reading assessment. *Assessment for Effective Intervention, 44*, 160-172. https://doi.org/10.1177/1534508418758371

Kilgus, S., Chafouleas, S. M., & Riley-Tillman, T. C. (2013). Development and initial validation of the social and academic behavior risk screener for elementary grades. *School Psychology Quarterly*, *28*(3), 210-226. https://doi.org/10.1037/spq0000024

Kim, G. Y., 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*, 1287–1310. https://doi.org/10.1007/s11145-017-9724-6

Lee, W.-C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*, 1-17. https://www.jstor.org/stable/25651533

Mercer, S. H., Keller-Margulis, M. A., Faith, E. L., Reid, E. K., & Ochs, S. (2018). The potential for automated text evaluation to improve the technical adequacy of written expression curriculum-based measurement. *Learning Disability Quarterly.* Advance online publication. https://doi.org/10.1177/0731948718803296

Mercer, S. H., Keller-Margulis, M. A., Faith, E. L., Reid, E. K., & Ochs, S. (2018). The potential for automated text evaluation to improve the technical adequacy of written expression curriculum-based measurement. *Learning Disability Quarterly, 42*, 117-128. https://doi.org/10.1177/0731948718803296

Mercer, S. H., Martinez, R. S., Faust, D., & Mitchell, R. R. (2012). Criterion-related validity of curriculum-based measurement in writing with narrative and expository prompts relative to passage copying speed in 10th grade students. *School Psychology Quarterly, 27*(2), 85–95. https://doi.org/10.1037/a0029123

National Assessment Governing Board (NAGB). (2017). *Writing framework for the 2017 National Assessment of Educational Progress*. Washington, DC: Author. Retrieved from https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/writing/2017-writing-framework.pdf

National Commission on Writing for America's Families, Schools, and Colleges. (2004). *Writing: A ticket to work…or a ticket out. A survey of business leaders.* Iowa City, IA: The College Board.

National Commission on Writing for America's Families, Schools, and Colleges. (2005). *Writing: A powerful message from state government.* Iowa City, IA: The College Board.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core state standards.* Washington, DC: Author. Retrieved from http://www.corestandards.org/

Northwest Regional Educational Laboratory. (2004). *An introduction to the 6+1 trait writing assessment model*. Portland, OR: Author.

Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology, 101,* 37–50. https://doi.org/10.1037/a0015586

Olinghouse, N. G., Graham, S., & Gillespie, A. (2015). The relationship of discourse and topic knowledge to fifth graders' writing performance. *Journal of Educational Psychology, 107*, 391-406. https://doi.org/10.1037/a0037549

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing, 26*, 45-65. https://doi.org/10.1007/s11145-012-9392-5

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis, & J. C. Burstein (Eds.). *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 19*, 139–158. https://doi.org/10.1080/10573560308222

Palermo, C., & Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology, 54*, 255-270. https://doi.org/10.1016/j.cedpsych.2018.07.002

Parker, R., Tindal, G., & Hasbrouck, J. (1991). Countable indices of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality, 2*, 1-17. https://doi.org/10.1080/09362839109524763

Pendergast, L. L., Youngstrom, E. A., Ruan-lu, L., & Beysolow, D. (2018). The nomogram: A decision-making tool for practitioners using multitiered systems of support. *School Psychology Review, 47*, 345-359. DOI: 10.17105/SPR-2017-0097.V47-4

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 121-132). Anderson, SC: Parlor Press.

Perelman, L. (2014). When 'the state of the art' is counting words. *Assessing Writing, 21*, 104-111. https://doi.org/10.1016/j.asw.2014.05.001

Persky, H. R., Daane, M. C., & Jin, Y. (2002). *The Nation's Report Card: Writing 2002. (NCES 2003-529).* Washington, D.C.: National Center for Education Statistics, Institute of Education Sciences. U.S. Department for Education.

Rodrigues, J., Jordan, N. C., & Hanson, N. (2019). Identifying fraction measures as screeners of mathematics risk status. *Journal of Learning Disabilities, 52*, 480-497. https://doi.org/10.1177/0022219419879684

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46*, 343-366. https://doi.org/10.1016/j.jsp.2007.06.006

Romig, J. E., Therrien, W. J., & Lloyd, J. W. (2017). Meta-analysis of criterion validity for curriculum-based measurement in written language. *Journal of Special Education,* 51, 72-82. https://doi.org/10.1177/0022466916670637

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, *105*, 1010-1025. https://doi.org/10.1037/a0032340

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, New York: John Wiley & Sons.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*, 1–30. https://doi.org/10.1191/0265532205lt295oa

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. Van Steendam, M. Tillema, G. Rijlaarsdam & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (Vol. 27, pp. 1–22). Leiden: Brill.

Schoonen, R. (2019). Are reading and writing building on the same skills? The relationship between reading and writing in L1 and EFL. *Reading and Writing*, *32*, 511-535. https://doi.org/10.1007/s11145-018-9874-1

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

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20,* 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62,* 5–18. https://doi.org/10.1177/0013164402062001001

Silberglitt, B., and 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. https://doi.org/10.1177/073428290502300402

Slomp, D. H. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing, 17,* 81-91. https://doi.org/10.1016/j.asw.2012.02.001

Smarter Balanced Assessment Consortium. (2017a). *Estimated testing times for Smarter Balanced summative assessments.* Retrieved from https://portal.smarterbalanced.org/library/en/estimated-testing-times.pdf

Smarter Balanced Assessment Consortium. (2017b). *Smarter Balanced cut score validation: Final report.* Retrieved from http://portal.smarterbalanced.org/library/en/smarter-balanced-cut-score-validation-final-report.pdf

Smarter Balanced Assessment Consortium. (2017c). *Smarter Balanced Assessment Consortium: 2015-16 summative technical report.* Retrieved from https://portal.smarterbalanced.org/library/en/2015-16-summative-technical-report.pdf

Smola, A. J., & Scholkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*, 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Smolkowski, K., & Cummings, K. D. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention, 41*, 41–54. https://doi.org/10.1177/1534508415590386

STARD. (2015). *An updated list of essential items for reporting diagnostic accuracy studies.* Retrieved from http://www.stard-statement.org/.

Stevenson, N. A. (2017). Comparing curriculum-based measures and extant datasets for universal screening in middle school reading. *Assessment for Effective Intervention, 42*, 195-208. https://doi.org/10.1177/1534508417690399

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51-65. https://doi.org/10.1016/j.asw.2013.11.007

Swets, J., & Picket, R. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory.* New York: Academic Press.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review, 42,* 402–414.

VanDerHeyden, A. M., & Burns, M. K. (2018). Improving decision making in school psychology: Making a difference in the lives of students, not just a prediction about their lives. *School Psychology Review, 47*, 385-395. DOI: 10.17105/SPR-2018-0042.V47-4

Van Norman, E. R., Klingbeil, D. A., & Nelson, P. M. (2017). Posttest probabilities: An empirical demonstration of their use in evaluating the performance of universal screening measures across settings. *School Psychology Review, 46*, 349-362. DOI: 10.17105/SPR-2017-0046.V46-4

White, S., Kim, Y., Chen, J., and Liu, F. (2015). *Performance of fourth-grade students in the 2012 NAEP computer-based writing pilot assessment: Scores, text length, and use of editing tools (NCES 2015-119).* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in Grades 3 and 4. *Journal of School Psychology, 68*, 19-37. https://doi.org/10.1016/j.jsp.2017.12.005

Wilson, J., & Chen, D., Sandbank, M. P., & Hebert, M. (2019). Generalizability of automated scores of writing quality in grades 3-5. *Journal of Educational Psychology, 111*, 619-640. https://doi.apa.org/doi/10.1037/edu0000311

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

Wilson, J., Olinghouse N. G., & Andrada, G. N. (2014). Does automated feedback improve writing quality? *Learning Disabilities: A Contemporary Journal, 12*, 93-118. https://eric.ed.gov/?id=EJ1039856

Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing, 27*, 11-23. https://doi.org/10.1016/j.asw.2015.06.003

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research, 58*, 87-125. https://doi.org/10.1177%2F0735633119830764 Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment, 2*, 37-56.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*, 561–577.

Table 1

Demographics of Study Participants.

	Grade 3 (*n* = 185)	Grade 4 (*n* = 167)	Grade 5 (*n* = 187)
Sex			
Male	76 (41%)	84 (50%)	85 (45%)
Female	109 (59%)	83 (50%)	102 (61%)
Race			
African American	13 (7%)	18 (11%)	14 (7%)
Asian	7 (4%)	14 (8%)	12 (6%)
Hispanic/Latino	43 (23%)	47 (28%)	51 (27%)
White	122 (66%)	88 (53%)	110 (59%)
Special Education	19 (10%)	26 (16%)	22 (12%)
English-language learner	36 (19%)	38 (23%)	42 (22%)
Chronological age[a]: *M* (*SD*)	104.31 (4.33)	115.78 (4.62)	128.33 (4.84)
WIAT Sentence Composition[b]: *M* (*SD*)	100.51 (14.58)	98.33 (15.53)	98.81 (14.41)
Base rate[c]: (%)	40%	40%	30%

Note. Reported values are frequencies.
[a]Chronological age reported in months. [b]Wechsler Individual Achievement Test 3rd Edition Sentence Composition subtest scored using age-based standard scores (*M* =100; *SD* = 15). [c]Base rate = the percent of students in the sample scoring in the at-risk range on the Spring 2016 Smarter Balanced ELA assessment (i.e., Levels 1 and 2).

Table 2

Correlation of Screener Scores with Risk Status and Descriptive Statistics for Grade 3.

	Risk Status	Descriptives					
		M	*SD*	Median	Range	Z-skewness	Z-kurtosis
Avg. of 6 prompts	-.47	10.22	2.65	10.67	6.00-18.17	0.54	-1.22
Avg. of 5 prompts	-.47	10.27	2.65	10.60	6.00-18.00	0.42	-1.34
Avg. of 4 prompts	-.46	10.26	2.66	10.75	6.00-18.00	-0.06	-1.99
Avg. of 3 prompts	-.45	10.33	2.71	11.00	6.00-18.00	-0.34	-1.96
Avg. of 2 prompts	-.45	10.41	2.89	11.00	6.00-18.00	-0.23	-1.62
First prompt	-.44	10.62	3.31	12.00	6.00-18.00	0.20	-1.42

Note. N = 185. Correlations with Risk status are presented in the first column. Risk status is a dummy coded variable (0 = not at risk; 1 = at risk) based on students' Spring Smarter Balanced ELA test level. All screeners scored using the PEG holistic quality score (range = 6-30). All correlations are statistically significant at $p < .001$.

Table 3

Correlation of Screener Scores with Risk Status and Descriptive Statistics for Grade 4.

	Risk Status	Descriptives					
		M	*SD*	Median	Range	Z-skewness	Z-kurtosis
Avg. of 6 prompts	-.48	11.63	2.59	11.83	6.00-17.67	-0.24	-0.11
Avg. of 5 prompts	-.48	11.68	2.68	12.00	6.00-17.40	-0.13	-0.33
Avg. of 4 prompts	-.48	11.78	2.77	12.00	6.00-18.00	0.22	-0.46
Avg. of 3 prompts	-.49	11.97	2.87	12.00	6.00-18.67	0.66	-0.07
Avg. of 2 prompts	-.44	12.04	3.03	12.00	6.00-19.50	0.77	-0.21
First prompt	-.42	12.25	3.33	12.00	6.00-19.00	0.19	-0.15

Note. N = 167. Correlations with Risk status are presented in the first column. Risk status is a dummy coded variable (0 = not at risk; 1 = at risk) based on students' Spring Smarter Balanced ELA test level. All screeners scored using the PEG holistic quality score (range = 6-30). All correlations are statistically significant at $p < .001$.

Table 4

Correlation of Screener Scores with Risk Status and Descriptive Statistics for Grade 5.

	Risk Status	Descriptives					
		M	*SD*	Median	Range	*Z*-skewness	*Z*-kurtosis
Avg. of 6 prompts	-.44	11.70	2.63	11.83	6.00-19.33	1.95	1.45
Avg. of 5 prompts	-.45	11.67	2.65	12.00	6.00-19.60	1.52	0.86
Avg. of 4 prompts	-.45	11.75	2.66	12.00	6.00-19.25	1.41	0.60
Avg. of 3 prompts	-.44	11.85	2.66	12.00	6.00-19.67	1.28	1.36
Avg. of 2 prompts	-.37	11.91	2.80	12.00	6.00-19.00	0.58	0.95
First prompt	-.33	12.16	3.05	12.00	6.00-18.00	-0.04	1.07

Note. N = 187. Correlations with Risk status are presented in the first column. Risk status is a dummy coded variable (0 = not at risk; 1 = at risk) based on students' Spring Smarter Balanced ELA test level. All screeners scored using the PEG holistic quality score (range = 6-30). All correlations are statistically significant at $p < .001$.

Table 5

Measures of Classification Accuracy for the Average PEG Score of Six Writing Prompts at Each Grade.

	Cutpoint	Sensitivity	Specificity	PP+	PP-	PP+ - BR	PP- - BR
Grade 3 (*n* = 185; base rate = 40%)							
d-based cutpoint	9.58	.66	.80	.69	.22	.29	-.18
90% Sensitivity-based cutpoint	11.75	.90	.49	.54	.12	.14	-.28
80% Sensitivity-based cutpoint	11.42	.84	.53	.54	.17	.14	-.23
Grade 4 (*n* = 167; base rate = 40%)							
d-based cutpoint	10.92	.64	.83	.72	.22	.32	-.18
90% Sensitivity-based cutpoint	12.42	.90	.47	.53	.13	.13	-.27
80% Sensitivity-based cutpoint	12.08	.84	.58	.57	.15	.17	-.25
Grade 5 (*n* = 187; base rate = 30%)							
d-based cutpoint	11.58	.80	.73	.56	.11	.16	-.29
90% Sensitivity-based cutpoint	12.42	.90	.47	.53	.13	.13	-.27
80% Sensitivity-based cutpoint	12.08	.84	.58	.57	.15	.17	-.25

Note. PEG Score range = 6-30. PP+ = positive posttest probability. PP- = negative posttest probability. PP+ - BR = the increase in probability of correctly identify a truly at risk student calculated by subtracting the base rate from the PP+. PP- - BR = the reduction in the probability of making a false negative decision calculated as subtracting the base rate from the PP-.

Table 6

Measures of Classification Accuracy for the PEG Score for the "Most Efficient" Writing Screener at Each Grade.

	Cutpoint	Sensitivity	Specificity	PP+	PP-	PP+ - BR	PP- - BR
Grade 3 (*n* = 185; base rate = 40%)							
Single Prompt Screener							
d-based cutpoint	11.50	.74	.75	.66	.19	.26	-.21
90% Sensitivity-based cutpoint	12.50	.91	.24	.45	.20	.05	-.20
Grade 4 (*n* = 167; base rate = 40%)							
Single Prompt Screener							
d-based cutpoint	11.50	.74	.75	.66	.19	.26	-.21
90% Sensitivity-based cutpoint	12.50	.91	.24	.45	.20	.05	-.20
Grade 5 (*n* = 187; base rate = 30%)							
Three Prompt Screener							
d-based cutpoint	11.83	.76	.73	.54	.12	.14	-.28
90% Sensitivity-based cutpoint	12.17	.91	.42	.40	.08	.00	-.32

Note. PEG Score range = 6-30. PP+ = positive posttest probability. PP- = negative posttest probability. PP+ - BR = the increase in probability of correctly identify a truly at risk student calculated by subtracting the base rate from the PP+. PP- - BR = the reduction in the probability of making a false negative decision calculated as subtracting the base rate from the PP-.

Table 7

Hanley-McNeil Comparisons of AUCs Generated from Writing Screeners Scored for AES and Word Count.

	PEG Score		Word Count		
	AUC (*SE*)	95% CI	AUC (*SE*)	95% CI	Comparison
Grade 3					
Six Prompt Screener	.78 (.034)	[.72, .85]	.69 (.042)	[.61, .78]	$Z = 3.49$, $p < .001$
Single Prompt Screener	.76 (.037)	[.69, .83]	.68 (.046)	[.59, .77]	$Z = 2.40$, $p = .016$
Grade 4					
Six Prompt Screener	.79 (.036)	[.72, .86]	.71 (.043)	[.62, .79]	$Z = 2.96$, $p = .003$
Single Prompt Screener	.76 (.038)	[.68, .83]	.68 (.045)	[.59, .77]	$Z = 2.42$, $p = .016$
Grade 5					
Six Prompt Screener	.81 (.035)	[.74, .87]	.71 (.042)	[.62, .79]	$Z = 4.61$, $p < .001$
Three Prompt Screener	.79 (.036)	[.72, .86]	.69 (.045)	[.57, .75]	$Z = 4.39$, $p < .001$

Note. AUC = area under the receiver operating characteristic curve.