Examining the Number of Writing Prompts Required to Reliably Evaluate Elementary Students' Writing Ability

Automated essay scoring (AES) systems are increasingly being used in classroom contexts because they help reduce teachers' grading load and accelerate the practice-feedback loop needed to improve students' writing skills. However, questions remain about the validity of score inferences from AES for specific instructional uses, such as obtaining reliable and accurate evaluations of students' writing ability, particularly that of struggling writers. This study explored this topic in the context of the PEG Writing AES system developed by Measurement Incorporated. We specifically sought to answer the following questions: How many prompts are required to obtain a generalizable estimate of elementary grade students' writing ability? Does the number of prompts differ based on if a student is classified as a struggling writer?

In the 2015-16 school year, 574 students in Grades 3-5 from 31 classrooms in three elementary schools, for whom consent and assent was obtained, completed a total of six prompts within 6 days. Students who scored at or below the 25th percentile on WIAT writing subtests (Sentence Combining and Sentence Building) and on a handwriting fluency task were identified as struggling writers (n = 101). Students who scored at or above the 30th percentile on those measures were regarded as not struggling (n =440).

All students completed two prompts in three genres—narrative, informative, and persuasive—totaling 6 prompts. Within genre, the prompts were randomly assigned (without replacement) from a set of six possible prompts per genre (18 prompts total). Students were given 30 minutes to plan and draft their responses, which were subsequently transcribed into Word documents and copied and pasted into PEG Writing for evaluation.

Using an analytic method called generalizability theory (i.e., G-theory), we found that PEG offers a feasible and efficient method to reliably assess students' "true" writing ability. For instance, the results indicated that two writing prompts per genre provided a reliable estimate of non-struggling writers' writing ability but struggling writers require a total of three prompts per genre. Results are encouraging with respect to results of prior G-studies that focused on writing prompts that were scored by humans. Graham et al. (2014) found that 11 prompts were necessary. Kim et al. (2017) found 6 prompts and 4 raters, or 7 prompts and 2 raters were necessary. Thus, results of the present study extend nascent research on the validity of AES use in classroom contexts, and have positive implications for the feasible evaluation of students' writing ability, particularly for struggling writers.