2016

Assessing the Writing Achievement of Young Struggling Writers: Application of Generalizability Theory

Steve Graham
Arizona State University, steve.graham@asu.edu

Michael Hebert
University of Nebraska-Lincoln, michael.hebert@unl.edu

Michael Paige Sandbank
Vanderbilt University

Karen R. Harris
Australian Catholic University

Learning Disability Quarterly 2016, Vol. 39(2) 72–82

DOI: 10.1177/0731948714555019

Assessing the Writing Achievement of Young Struggling Writers: Application of Generalizability Theory

Steve Graham, EdD[1,2], Michael Hebert, PhD[3], Michael Paige Sandbank, MA[4], and Karen R. Harris, EdD[1,2]

1. Australian Catholic University, Brisbane
2. Arizona State University, Tempe, USA
3. University of Nebraska–Lincoln, USA
4. Vanderbilt University, Nashville, TN, USA

Corresponding Author:
Steve Graham, Arizona State University, Mary Lou Fulton Teachers College, P.O. Box 871811, Tempe, AZ, USA. Email: Steve.graham@asu.edu

Abstract
This study examined the number of writing samples needed to obtain a reliable estimate of young struggling writers' capabilities. It further assessed if performance in one genre was reflective of performance in other genres for these children. Second- and third-grade students (81 boys, 56 girls), who were identified as struggling writers in need of special assistance by their teacher and scored at the 25th percentile or lower on a norm-referenced story-writing test, wrote four compositions: a story, personal narrative, opinion essay, and informative text. Applying generalizability theory (G-theory), students' scores on three writing measures (total number of words [TNW], vocabulary diversity, and writing quality) for the four compositions were each portioned into variance due to the following sources: students, writing tasks, and the interaction between students and writing tasks. We found that 14, 8, and 11 compositions, respectively, would be needed to obtain a reliable estimate of these students' writing capabilities in terms of TNW, vocabulary diversity, and writing quality. Furthermore, how well these students wrote in one genre provided a weak prediction of how well they wrote in other genres.

Keywords
at risk, writing, assessment

Identifying Students in Need of Special Writing Instruction

Writing is the primary means by which students demonstrate their knowledge in today's classrooms (Graham, 2006). Students use writing to gather and organize knowledge, explore and refine their ideas, and show what they know (Durst & Newell, 1989). Writing about text read and teaching writing also have a positive impact on reading outcomes (Graham & Hebert, 2011), whereas writing about material presented in class enhances learning (Bangert-Drowns, Hurley, & Wilkinson, 2004). Failure to acquire strong writing skills, therefore, restricts opportunities for postsecondary education and employment. For

instance, employers report they rely on writing when making decisions about who to hire and promote (National Commission on Writing, 2004, 2005).

Despite the importance of writing, many children experience difficulty learning how to write. Data from the National Center for Education Statistics (2012) reveal that less than a third of students in the United States have mastered the skills necessary for *proficient*, or grade-level appropriate writing. The vast majority of youngsters in the United States scored at the *Basic* level or below, which denotes only partial mastery of the writing skills needed at each grade
(Loomis & Bourque, 2001). Clearly, it is important to identify students who experience difficulty with writing early, so they can receive extra assistance or special instruction learning how to write. Waiting until later grades to address literacy problems that begin in the primary grades is often unsuccessful (Slavin, Madden, & Karweit, 1989). It is also more successful to address difficulties early, before they progress to a more severe level (Lyon, 1996).

Teachers play a critical role in identifying young students who might benefit from special instruction. They are often the first ones to recognize that a serious writing problem exists, referring children for assessment who they believe will benefit from special help. These referrals are followed by an assessment, typically involving a single writing sample from a norm-referenced test, to confirm or disconfirm the teacher's appraisal (e.g., Harris et al., 2012). The writing assessment commonly involves story writing as this is what is tested by most norm-referenced standardized writing tests (see review by Calfee & Wilson, 2004).

One assumption underlying this approach to identification is that one sample of writing is adequate for identifying students in need of special writing instruction. A second assumption is that performance in one genre is reflective of performance in other genres (e.g., when students write a story, it is assumed that the story composition provides a good index of their writing in general, including their writing in other genres such as opinion, personal narrative, and informative texts). Of course, this second assumption is not relevant if a teacher is only making a judgment about a stu- dent's competence in a single genre when they indicate a student would benefit from special writing instruction. While we are sure this happens, we think that teachers' judgments about which students need special writing instruction are typically based on a general judgment of writing competence rather than a genre-specific one.

Purpose of the Present Study

The present study tested the veracity of the two assumptions above by asking the following questions:

1. How many writing samples are needed to obtain a reliable estimate of struggling students' writing achievement?
2. Does performance in one genre reflect performance in other genres?

Past studies conducted mostly with typically developing writers have shown that multiple samples of students' writing are needed to establish a reliable estimate of writing. For instance, in his seminal study published in 1966, Coffman found that a minimum of five writing tasks across genres were needed to reliably assess the writing achievement of typically developing students. In a more recent investigation, Huang (2008) reported that

three writing tasks were needed to obtain a reliable measure of writing ability for typically developing students, but five writing tasks were required to obtain a reliable estimate for English language learners in the same grades. Given these findings, it seems unlikely that a single sample of writing is adequate for confirming teachers' judgments that a serious writing problem exists.

It is also unlikely that students' performance in one genre provides a reliable estimate of writing in other genres. In two recent reviews of the writing assessment literature for typically developing writers, Graham and colleagues (Graham, Harris, & Hebert, 2011a; Graham, Hebert, & Harris, 2011) found (a) quality of students' writing differed meaningfully from one genre to the next and (b) quality of
students' writing in one genre was a weak predictor of their performance in other genres. More specifically, students' mean writing scores differed statistically across genres, and correlations for students from one genre to the next were small to moderate (.10 to .60).

Why was performance in one genre not reflective of performance in other genres in the studies reviewed by Graham and colleagues (Graham et al., 2011a; Graham et al., 2011)? The most obvious answer to this question is genres differ in terms of rhetorical structures, basic elements, and even the types of words students use (Donovan & Smolkin, 2006). Typically developing writers are more knowledgeable about some writing genres than others (see Gillespie, Olinghouse, & Graham, 2013; Klein & Rose, 2010; Lin, Monroe, & Troia, 2007), and this genre knowledge shapes how young children write. They use this knowledge to determine the basic structure of their text, and it guides their search for relevant writing content and words when composing (Bereiter & Scardamalia, 1987; Hayes, 2011). In other words, writing a story is different than writing an opinion essay, as the construction of each draws on different types of knowledge and abilities. Consequently, children's scores for these two types of writing need not converge, especially because a child's knowledge about one genre may be more or less complete than knowledge about another genre.

This line of reasoning, however, may not be valid for young children who struggle with writing. They know much less about writing than their typically developing peers (Englert, Raphael, Fear, & Anderson, 1988; Saddler & Graham, 2007), and it is possible they make few distinctions between different genres (Lin et al., 2007). Just as importantly, they may not apply the knowledge they possess as Saddler and Graham (2007) found that these children's general knowledge about writing was not statistically related to their writing performance.

Generalizability Theory (G-Theory)

To answer our first question (How many writing samples are needed to obtain a reliable estimate of struggling students' writing achievement?), we applied G-theory. G-theory, developed by Cronbach (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is an extension of classical testing theory, and provides a method for determining ". . . how much of a students' score is attributable to actual capability (true score) and how much to error (factors unrelated to capability)" (Novak, Herman, & Gearhart, 1996).

In our study, second- and third-grade students who were identified by their teachers as struggling writers and scored at or below the 25th percentile on a norm-referenced story- writing test wrote four compositions: a story, personal narrative, an opinion essay, and an informative text. Each of these types of writing is emphasized by the Common Core State Standards (2010). All four compositions were scored for total number of words

(TNW), vocabulary diversity, and writing quality. Applied to data such as ours, G-theory partitions total score variance into variance due to the object of measurement (students' scores on the writing samples), total variance due to the object and conditions of measurement (in this study, multiple writing tasks), and the variance from the interaction. This portioning of variance allowed us to compute generalizability (G) coefficients that provided an estimate of the reliability of students' writing achievement (or true score) for the three writing measures (TNW, vocabulary diversity, and writing quality) by the number of writing samples administered (one to four).

While Brown, Glasswell, and Harland (2004) indicated that coefficients exceeding .80 are a robust indication that multiple writing tasks are tapping a common construct (i.e., writing achievement), we applied a more stringent criteria of .90 for the coefficient. The identification of writers who are in need of special services involves making decisions about individual students. As Nunnally (1967) argued, a coefficient below .90 is not acceptable for this purpose. Because the four writing samples completed by the young struggling writers in this study did not reach the .90 criteria, we also conducted a Decision study (D-study; Shavelson & Webb, 1991). This involves using the data from the G-study to estimate how many writing samples are needed to obtain an acceptable level of reliability.

To the best of our knowledge, researchers have not examined how many different writing samples are needed to obtain a reliable estimate of writing for students who are potential candidates for special writing instruction. The answer to this question is important not only in identifying students who should receive such instruction but in con- ducting research and interpreting the results of high-stakes and classroom writing assessments with this group of students. As Schoonen (2005) noted, ". . . in research and assessment, different writing assignments are generally considered to be samples of the same 'universe of admissible observations' and used to generalize about a person's writing proficiency" (p. 2).

To answer our second question (Does performance in one genre reflect performance in other genres?), we examined if students' performance on the four different genres of writing differed statistically one from the other and whether performance on these tasks were statistically correlated. If struggling writers possess and apply varying amounts of genre knowledge when writing, it is reasonable to anticipate differences in students' writing across genres for our three writing measures (i.e., TNW, vocabulary diversity, and writing quality). For example, students who possess greater knowledge of a particular genre should be able to use this knowledge to define the task, retrieve relevant information (including vocabulary), and verify the appropriateness of the retrieved ideas, resulting in higher quality compositions with more ideas and more diverse vocabulary (Bereiter & Scardamalia, 1987; Hayes, 2011). We further expected that these anticipated differences in writing performance across genres would be reflected in modest correlations between the four different types of writing.

Method

Participants

Twelve third-grade and 11 second-grade teachers in four schools in a single urban school district in Washington, D.C., were asked to consider the 590 students in their class- rooms and identify those who they believed were struggling writers who would benefit from special instruction in learning how to write. The identified students were administered Form B of the story construction subtest of the *Test of Written Language*–3 (TOWL-3; Hammill & Larsen, 1996). This subtest assesses a child's ability to write an interesting and complete

story including thematic elements such as plot, character development, and general composition. A graduate student scored the TOWL-3 after identifying information had been removed. A second graduate student unfamiliar with the design and purpose of the study rescored half of the tests. Interrater reliability was .80.

All told, 156 students identified by teachers as struggling writers needing special writing instruction scored at or below the 25th percentile on this test (see Note 1). Although no true cut point exists for risk in writing, we selected the 25th percentile as the cut point in this study as young students scoring at this level or below have benefited for specialized writing instruction (e.g., Graham, Harris, & Fink-Chorzempa, 2002; Harris, Graham, & Mason, 2006).

The parents of 138 students provided informed consent for their children to participate in our study. However, one third-grade student was dropped from the analysis due to missing data. The final 137 participants included 63 second-grade students and 74 third-grade students. Eighty-one students were boys and 56 were girls (see Table 1). The mean age of the participants was 8 years 0 months (SD = 0.63). The majority of the students in the sample were identified as Black (72%), with smaller percentages of students identified as White (13%), Asian (9%), and Hispanic (6%). This was consistent with the demographics of the participating urban school district. The number of students identified as receiving free or reduced lunch was 61% of the sample. A total of 32 students were identified as having a disability: 15 were identified as having a learning disability, 11 with speech and language impairments, 4 with emotional/behavioral disorders, and 2 with attention-deficit/ hyperactivity disorder (ADHD).

Table 1. Student Characteristics by Grade Level.

	Grade level	
Variable	Second	Third
Sample size	63	
Age (months)		
M (*SD*)	89.25 (4.73)	101.07 (4.87)
Gender		
Male	37	44
Female	26	30
Race		
Black	44	55
White	8	10
Asian	4	8
Hispanic	7	1
Free or reduced lunch		
Yes	36	50
No	27	24
Students with disabilities		
Yes	13	20
No	50	52
Unknown	0	2

Setting

Prior to the start of the study, teachers in all 23 classrooms were interviewed to determine their approach to writing instruction. All teachers indicated they used a Writer's Workshop model (Calkins, 1986; Graves, 1983). Next, they completed a survey on their writing practices (Graham, Harris, Fink-Chorzempa, & MacArthur, 2003) and their writing class

was observed by graduate assistants three or four times to verify and enhance this initial description. The observers took informal notes during their observations, recording all of the instructional activities they witnessed.

The data collected from these two sources confirmed that teachers did in fact apply a Workshop approach. All teachers established a writing routine where students were expected to plan their composition, write a first draft, revise and edit it, and publish the completed paper. Students were also expected to write frequently and for different purposes, with story and opinion writing receiving considerable emphasis. Students also wrote descriptions, personal narratives, poems, and book reports. Most of the teachers had students select their own writing topics and work at their own pace. Students conferenced with the teacher and peers about their writing, and they shared completed work and work in progress with the class. Despite these common themes, teachers differed in how they actualized Writers' Workshop, with some variation around core concepts such as how frequently students' shared their writing, selected their own writing topics, and were allowed to complete writing assignments at their own pace.

While teachers did not emphasize the teaching of writing skills during the initial interview, the survey and observations indicated that this was an important part of their writing classes. They primarily relied on mini-lessons several times a week to teach a variety of grammar and spelling skills. There was, however, considerable variability across teachers in terms of what was taught, including how much time they spent teaching handwriting, spelling, and grammar skills. Finally, some teachers taught students strategies for planning their compositions, mostly relying on strategies such as brainstorming, webbing, and Venn diagrams.

Collecting Samples of Students' Writing

The writing skills of students in the sample were assessed across four genres: story, personal narrative, opinion, and informative essay. Prior to the start of the study, eight writing prompts for each of these genres were evaluated for suitability of use by three primary-grade teachers as well as three students (two third-grade students and one second- grade student). Both teachers and students were given several prompts and asked to choose which prompts students in Grades 2 and 3 would (a) enjoy writing about and (b) be able to write about. The prompts used in this study were selected by all of the teachers and students.

To increase motivation for writing, students were given a choice of two prompts for each genre. For example, in story writing, students were asked to write a story in response to one of two picture prompts (i.e., two dogs rowing a boat *or* children baking something in the kitchen). Prompts in other genres were given in statement rather than picture form. For opinion writing, students were asked to state and defend their opinion on a home issue (i.e., Should children your age be allowed to choose their own pets? Should parents make their children clean their bedrooms?). For personal narra- tive writing, students were asked to write about something that happened to them (i.e., when they were younger *or* on the playground). For informative writing, students were asked to describe something (i.e., their favorite place *or* a family member). The two prompts for each genre produced equivalent writing performance for writing quality with second- and third-grade students in Graham, Harris, and Mason (2005) and Harris et al. (2006).

The four genre prompts were administered in a counter- balanced order across students. Students were assessed individually and provided as much time as they needed to complete their papers. They were assessed in only one genre on a single day to minimize

writing fatigue. Administrators informed the students that they could pro- vide help only with spelling, as we thought that they might produce more text if such help was provided. While the frequency of spelling requests was not recorded, examiners reported that such requests occurred only once or twice for most students. Once a student completed the writing assignment, he or she was asked to read it back to the examiner. This allowed us to identify any words that were illegible or so incorrectly spelled that they could not be identified.

As a group, students took about 20 min to write a personal narrative, opinion, or informative essay. They spent approximately 25 min writing stories.

Scoring Students' Writing

All student responses to the writing prompts for each genre were analyzed using the following three measures. The procedures for collecting data for the measures were employed consistently across genres and grade levels to ensure comparability of results.

TNW. Papers were scored for TNW. Every group of letters representing a spoken word was included in calculating TNW, regardless of spelling. All papers were scored by a rater unfamiliar with the purpose of the study, with 50% of the papers independently scored by a second rater unfamiliar with the purpose of the study. Interrater reliability was .99 for both grades.

Vocabulary diversity. Papers were further scored using the formula for *Corrected Type-Token Ratio* (CTTR): Number of different words/ 2×TNW . This is a measure of vocabulary diversity, corrected for the length of composition (Hess, Haug, & Landry, 1989). Number of words was determined by subtracting each repetition of a word in the composition from the number of total words. All papers were scored for CTTR by a graduate student rater, with 50% of the papers scored by a second rater. Interrater reliability for CTTR was .98 for the second-grade sample and .96 for the third-grade sample.

Writing quality. Papers were also scored for writing quality using a holistic scale (Diederich, 1966). Scores ranged from 1 to 8, with 1 representing the lowest score possible and 8 representing the highest. Raters were asked to read through each student composition carefully to obtain an overall impression of the writing quality. They were further instructed to consider writing quality as a combination of ideation, organization, sentence structure, grammar, and word choice with no single factor receiving undue weight.

Raters were provided with three anchor papers, representative of a low, middle, and high paper for each genre. The exemplar papers were obtained in participating schools from students in second- and third-grade classrooms that did not participate in this study. All children in these class- rooms wrote responses to the prompts used in the study. From these students' papers, two former elementary school teachers selected the best, average, and poorest quality compositions from each genre to serve as anchor papers for scoring. In selecting the anchor points, the two former elementary school teachers were asked to focus equally on ideation, organization, sentence structure, grammar, and word choice when considering the overall quality of a composition.

All papers written by students participating in our investigation were typed and corrected for spelling, capitalization, and punctuation prior to scoring for writing quality. This was done to minimize possible bias due to examiners placing undue weight on surface features of the responses. Previous research has shown that features such as handwrit-

ing, spelling, and usage miscues can have excessive influence over judgments of writing quality (Graham, Harris, & Hebert, 2011b).

Writing quality was scored by former primary-grade teachers unfamiliar with the purpose of the study. Examiners were trained how to use the 8-point scales and anchor points when scoring each type of writing (stories, personal narrative, opinion, and informative). For each genre, the writing quality score was the average score of the two raters. Interrater reliabilities for story, opinion, personal narrative, and informational writing quality ratings were .91, .91, .89, and .79, respectively, for the second-grade sample, and .87, .93, .89, and .72, respectively, for the third-grade sample.

Table 2. Means and Standard Deviations of Writing Measures Collected for Each Genre.

	Genre			
	Story	Personal narrative	Opinion	Informative
Measure	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
TNW	40.27 (29.83)	36.48 (29.05)	29.02 (25.90)	38.03 (31.43)
Vocabulary	2.80 (0.69)	2.60 (0.82)	2.53 (0.81)	2.85 (1.56)
Writing quality	2.37 (1.69)	2.15 (1.51)	2.11 (1.42)	2.32 (1.67)

Note. TNW = total number of words; Vocabulary = vocabulary diversity.

Table 3. Mean Square Error (MSE) and Percentage of Variance for Each Source of Variance for Each Writing Measure.

	Mean square error and percent of variance for each source of variance		
Measure	Students	Writing Tasks	Students × Writing Tasks
TNW			
MSE	1,834.82	3,254.55	519.12
Proportion of variance	37.9%	2.3%	59.8%
Vocabulary diversity			
MSE	1.68	1.89	.29
Proportion of variance	53%	1.8%	45.2%
Writing quality			
MSE	5.86	2.20	1.33
Proportion of Variance	45.9%	0.3%	53.8%

Note. TNW = total number of words.

Results

Data Modifications

As noted earlier, one student was dropped from the study because of missing data (he did not complete several of the writing assessments). Three more students missed one writing assessment. To include these three students in the study, the PROC MI multiple imputation procedure in SAS software version 8.1 was used to impute these missing data. A Markov Chain Monte Carlo (MCMC) method was employed, which attempts to generate missing values from multidimensional probability distributions. The MCMC method uses maximum likelihood estimation to generate initial values for iterations of missing values, and then builds a stationary distribution from which the imputations are made (Yuan, 1990). The default of five imputations for this function was chosen, creating five complete data sets. The five data sets were then analyzed by the program using standard statistical analy-

ses, and the pro- gram averaged the results of the analyses from the complete-data estimates for the inference. Missing data were imputed using data from all writing measures, as well as demographic data including grade level, gender, and race. While the MCMC method assumes multivariate normality, the inferences made are robust if the amounts of missing data are not large (Yuan, 1990), which was the case in this study.

It was further necessary to transform scores for the writing quality variables prior to conducting parametric statistical analyses, as they were not normally distributed. A square root transformation was applied. Lack of normality in the writing quality scores was not surprising, as students were selected on the basis of low writing performance. Transforming the writing quality scores was not viewed as problematic, because the original scale for writing quality was arbitrary. In other words, the researchers defined the scale to be used, and altering the scale did not affect the interpretation of the models.

As noted earlier, students were given a choice of two topics to write about for each genre. This was done to increase motivation. In this study, TNW, vocabulary diversity, and writing quality did not differ statistically when students who selected one of the two topics in a genre were compared with students who wrote about the other topic (all *p*s > .29). Thus, writing topic within a genre was not used as a factor in any of the analyses.

> **Question 1:** How many writing samples are needed to obtain a reliable estimate of struggling students' writing achievement?

To determine how many different writing samples need to be collected to obtain a reliable estimate of the three writing measures (i.e., TNW, vocabulary diversity, and writing quality), G-theory methodology was applied. For our study, G-theory was used to partition total score variance into variance due to the object of measurement (students' scores on the writing samples), total variance due to the object and conditions of measurement (in this study, multiple writing tasks), and the variance from the interaction. This portioning of variance allowed us to compute generalizability (G) coefficients. These provided estimates of the reliability of students' writing achievement (or true score) for the three writing measures used to score the four writing tasks. This approach further allowed us to conduct a D-study to ascertain the number of tasks needed to obtain an acceptable level of reliability (.90). Means and standard deviations for each writing measure by genre are presented in Table 2 (scores for writing quality in Table 2 are not transformed).

In this study, we employed a fully crossed one-facet design with three sources of variance for each measure: (a) Students, (b) Writing Task, and (c) Students × Writing Tasks interaction. Scores on the writing task served as the dependent variable. Students were considered the object of measurement and the Students × Writing Tasks interaction the source of error variance. For each writing measure, an analysis of variance (ANOVA) was conducted to partition variance attributable to each source. The estimates of variance from the ANOVAs were used to calculate G coefficients, as well as conduct a D-study to determine the effects of the number of writing tasks on the G coefficients. Mean square error and variance estimates for Students, Writing Tasks, and the interaction for the three writing measures are presented in Table 3.

These analyses allowed us to compute G coefficients that are appropriate for absolute decisions (used to categorize students into specific groups) or a relative decision (used to rank order students). Because we were most concerned with teachers' efforts to identify students with writing difficulties who would benefit from specialized writing instruction, we only reported G coefficients for absolute decisions (see Table 4). Such coefficients are

Table 4. Generalizability Coefficients for Absolute Decisions by Number of Writing Tasks.

Tasks	1	2	3	4	5	6	7	8	9	10	11	12	13	14
TNW	.38	.55	.65	.71	.75	.79	.81	.83	.86	.87	.88	.88	.89	.90
Vocab	.53	.69	.77	.82	.85	.87	.89	.90	.91	.92	.93	.93	.94	.94
Quality	.46	.63	.72	.77	.81	.84	.86	.87	.88	.89	.90	.911	.92	.922

Note. TNW = total number of words; Vocab = vocabulary diversity.

also relevant to high-stakes writing tests administered by the national government, states, and school districts that categorize students' performance (e.g., proficient/not proficient; see Note 2).

The G coefficients for absolute decisions appear in Table 4 in the column labeled "4" under the heading for Tasks, as this is the number of writing tasks used in this study. The coefficients from the D-study appear under the other columns for the heading Tasks. As can be seen in Table 4, the administration of the four writing tasks yielded G coefficients for absolute decisions below our criterion of .90 for all three measures. Consequently, the administration of the four different writing tasks did not yield a measure of writing achievement reliable enough to make individual decisions about students' writing. If we applied the less stringent criteria of .80 recommended by Brown et al. (2004), only vocabulary diversity was reliable enough to be used for this purpose.

The results of the D-study revealed that 14 writing tasks would need to be administered to obtain a G coefficient for absolute decisions of .90 for TNW. Eight writing tasks would be required to obtain this level of reliability for vocabulary diversity and 11 tasks for writing quality. With the less stringent criteria of .80, 7 writing tasks would be required for TNW, 4 for vocabulary diversity, and 5 for writing quality. The same number of tasks is needed to reliably rank order students' writing achievement (G coefficients for relative decisions) for both the .90 and .80 criteria.

Question 2: Does performance in one genre reflect performance in other genres?

A 2 (Grade) × 4 (Writing Tasks) ANOVA with repeated measures design was used for each of the three measures scored for each genre (i.e., TNW, vocabulary diversity, and writing quality). The homogeneity of variance assumption was met for each analysis, but the sphericity assumption was not met for any of them. To correct for this, the Greenhouse-Geisser correction was applied if the Greenhouse-Geisser test for sphericity was below .75 (this represents a conservative test); otherwise, the Huynd-Feldt correction was used because the Greenhouse-Geisser often fails to reject the null hypothesis when the estimate is greater than .75 (Field, 2000). When statistically significant effects were found for an ANOVA, post hoc analyses were conducted for genre comparisons. The p value was set at .0167 for the three ANOVAs (to control for Type 1 errors); post hoc analyses were set at the conventional p value of .05.

Comparison of TNW across genres. For the ANOVA involving TNW, the Huynd-Feldt correction was used because the Greenhouse-Geisser test indicated the probability of sphericity was .94. The only statistically significant result was for the main effect of genre, F(2.91,

392.3) = 6.09, p = .001. Post hoc analyses revealed students wrote fewer words when writing an opinion than when writing a story (p = .001) or informative essays (p = .005). No statistically significant differences were found for any other genre comparisons.

Comparison of vocabulary diversity across genres. For vocabulary diversity, the Greenhouse-Geisser correction was used because the indicated probability of sphericity was .62. The only statistically significant ANOVA result was for the main effect of genre, F(1.85, 244.09) = 6.60, p = .002. Post hoc analyses revealed students used fewer different words when writing an opinion than when writing a story (p = .002) or informative essay (p = .005). No other statistically significant differences were found between other genre comparisons.

Table 5. Correlations Between Writing Measure Variables for Each Genre.

	Genres					
Measure	S & PN	S & O	S & I	PN & O	PN & I	O & I
TNW	.30	.25	.37	.53	.48	.42
Vocabulary	.38	.39	.30	.49	.22	.37
Writing quality	.50	.45	.44	.60	.58	.52

Note. S = story; PN = personal narrative; O = opinion; I = informative; TNW = total number of words; Vocabulary = vocabulary diversity. All correlations were significant at the $p < .01$ level, two-tailed.

Comparison of writing quality across genres. For the ANOVA involving writing quality, the Huynd-Feldt correction was used because the Greenhouse-Geisser test indicated the probability of sphericity was .92. A statistically significant effect for genre was not obtained, F(2.86, 385.64) = 1.649, p = .180, but there was a statistically significant interaction between genre and grade, F(2.86, 385.64) = 7.37, $p < .001$. As a result, one-way ANOVAs with repeated measures were conducted separately for the second-grade and third- grade groups.

For the second-graders, the sphericity assumption was met (p = .126). Statistically significant main effects were found for genre, F(3, 186) = 4.49, p = .005. Post hoc analyses revealed students' opinion essays were of lower writing quality than their personal narratives (p = .028) and their informative essays (p = .004). No other statistically significant differences were found between genres at this grade.

For the third graders, the sphericity assumption for ANOVA was not met (p = .009). The Huynd-Feldt correction was used because the Greenhouse-Geisser test indicated the probability of sphericity was .89. A statistically significant main effect for genre was obtained, F(2.78, 203.19) = 3.50, p = .019. Post hoc analyses revealed students' opinion papers were of higher writing quality than personal narrative essays (p = .002) and informative essays (p = .005). No other statistically significant differences were found at this grade.

Correlations between the four different genres of writing. To determine if writing performance in one genre was related to writing performance in another genre, Pearson correla-

tion coefficients were calculated for each measure across genres (see Table 5). All correlations were statistically significant.

Correlations for TNW across genres were small to moderate in size ranging from .30 to .53. The smallest amount of variance was shared between story and opinion writing (6%), whereas the largest amount of variance was shared between personal narrative and opinion writing (28%).

Correlations for vocabulary diversity were also small to moderate, ranging between .22 and .49. Personal narrative and informative writing only shared 5% of variance for this measure, whereas personal narrative and opinion shared 24% of the variance in scores.

Correlations for writing quality were moderate in size ranging from .44 between story and informative writing, to .60 between personal narrative and opinion writing. Even so, shared variance between genres was only 18% and 36%, respectively.

Discussion

Teachers play a critical role in identifying young students who might benefit from special instruction to write. They often are the first ones to identify such students, referring them for special education services. The administration of these services requires confirmation that the teacher's appraisal was accurate, and this is commonly accomplished by administering a single writing assessment that evaluates how well a child writes in a single genre, such as story writing (Harris et al., 2012). The assumption underlying this approach is that a single composition provides a reliable estimate of students' writing achievement and that performance in one genre adequately reflects performance in other genres.

How Many Writing Samples Are Needed to Obtain a Reliable Estimate of Achievement?

The present study provides the first test of how many different writing samples are needed to obtain a reliable estimate of writing achievement for young students who are potential candidates for special writing instruction. While assessing writing using a single test may be parsimonious, it does not provide a reliable estimate of writing achievement for these children. A single composition yielded coefficients ranging from .38 to .53 for writing quality, TNW, and vocabulary diversity, respectively, for second- and third- grade students who were identified by their teachers as struggling writers in need of special writing instruction and who obtained low scores on a single assessment of their writing. This outcome mirrored findings with typically developing writers and English language learners (e.g., Coffman, 1966; Huang, 2008), indicating that students' writing achievements should not be based on a single composition.

Even when we included all four compositions (i.e., story, personal narrative, opinion, and informative) tested in this study, the resulting G coefficients for TNW, vocabulary diversity, and writing quality did not rise to our established benchmark of .90. For our most important measure, writing quality, reaching this benchmark required 11 different com- positions. If we set a less stringent level of .80, a criterion accepted as a robust indication that multiple writing tasks are tapping a common construct (Brown et al., 2004), only 5 compositions were required.

Findings from the generalizability analyses further indicated that individual differences among the participating students' writing performance contributed considerably to variance in students' writing scores, accounting for 38%, 53%, and 46% of variance in TNW, vocabulary diversity, and writing quality, respectively. This means students ac-

counted for 38% to 53% of variance in writing scores regardless of the type of writing collected. However, for writing quality and TNW, the greatest amount of variance in scores was attributable to the interaction between students and writing tasks, accounting for 54% and 60% of the variance, respectively. For vocabulary diversity, this interaction accounted for 45% of the variance in scores. If writing genre were irrelevant in assessing writing achievement, almost all of the variance would have been attributed just to students. In other words, little variance would have been attributed to the interaction between writing tasks and students.

Does Performance in One Genre Reflect Performance in Other Genres?

The present study is the first to demonstrate that young struggling writers' performance in one genre is not a reliable indicator of their performance in other genres. As predicted, how much students wrote, diversity in word choice, and overall quality of writing in one genre were not predictive of performance in the other genres for the young struggling writers in this study. The amount of common variance between two genres of writing for any measure never exceeded 36%. More typically, shared variance between genres on a specific measure ranged from 5% to 17%. Also as predicted, we found that mean performance on the three different writing measures (TNW, vocabulary diversity, and writing quality) was not identical for all four writing tasks. Specifically, students' opinion essays differed from their stories, personal narratives, and informative papers on one or more of these measures. While we anticipated differences between the other three genres as well, the collective findings from this study are consistent with research con- ducted with typically developing writers showing that children do better on some writing tasks than others, and writing in one genre is a poor predictor of performance in other genres (e.g., Graham et al., 2011a; Graham et al., 2011).

Conclusion

The assumption that a single composition provides a reliable estimate of the writing achievement of young students identified by teachers as needing special writing instruction is not valid based on the findings from the current study. Furthermore, it should not be assumed that how well these children write in one genre provides a strong reflection of their writing in other genres. These findings have obvious implications for identifying students in need of special instruction in learning to write. Multiple writing samples across different genres need to be collected to establish a reliable estimate of these students' writing achievement. This is no different from what we already know about assessing typically developing students' writing achievement (e.g., Coffman, 1966; Huang, 2008).

It is important to note that the findings from our study do not mean that 11 different writing tasks must be administered to obtain a reliable estimate of the quality of students' writing (this would not be cost-effective for schools). Not surprisingly, the number of writing tasks needed to establish reliable estimates varied from one sample of students to the
next in different studies (see, for example, Gearhart, Herman, Novak, & Wolf, 1995; Huang, 2008). More important, writing tasks are not the only facets that contribute to variance in writing scores. For example, fewer compositions are needed to establish a reliable estimate of writing achievement if the number of people scoring each composition is increased (Baker, Abedi, Linn, & Niemi, 1996; Swartz et al., 1999).

The implications of this study are not limited to the identification of students in

need of special writing instruction. Assessments are an essential element of schooling in the United States and elsewhere. They are used to measure students' accomplishments, determine what needs to be taught, gauge the effectiveness of instructional practices, and detect what should be modified. A fundamental assumption under- lying each of these purposes for struggling writers is that the assessments used are valid and reliable. In terms of many statewide high-stakes assessments, only a single sample of writing is typically collected in a school year (Calfee & Wilson, 2004; Jeffery, 2009). As this study demonstrated, it is unlikely that this is adequate for determining the writing achievement of students who teachers believe experience difficulty learning to write. Likewise, our findings indicate that teachers should be cautious when assessing students' writing, drawing on multiple samples of these children's writing when making decisions for classroom practices and instruction. Finally, it must be noted that a single sample of writing may not provide a reliable estimate of students' writing in a research study conducted with struggling writers.

As with all studies, the current investigation has several limitations. One, the study was conducted in an urban school system with second- and third-grade children. Most of the students were Black and lived in low socioeconomic status (SES) households. Consequently, additional research is needed to replicate this study in suburban and rural locations and with children beyond second and third grade.

Two, it must be recognized that the writing tasks students completed in the current study were not representative of all types of writing or the various purposes to which the four selected tasks can be applied. Furthermore, our writing outcomes were limited to TNW, vocabulary diversity, and writing quality (measured holistically). Thus, additional research is needed to determine if similar findings are obtained with other types of writing tasks (e.g., summary writing, cause and effect, problem solution) as well as writing measures (e.g., holistic vs. analytic writing measures).

We purposefully focused our efforts on determining how many different types of writing were needed to establish a reliable estimate of writing achievement. We could have taken a different approach and concentrated just on a single genre, such as informative writing. Hopefully, researchers in the future will do just that, examining how many writing samples are needed to provide a reliable estimate of writing competence in specific genres.

Last, there are many sources of variability that can con- tribute to variance in writing scores. In addition to type or genre of writing, possible facets that may contribute to variability include mode of writing (paper and pen or word processing), time constraints (timed or untimed), testing conditions (group or individual), and number of raters (one or more), to name but a few. Future research needs to examine the role of these possible sources of variance in the reliable assessment of struggling writers' accomplishments.

Declaration of Conflicting Interests : The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding : The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Unfortunately, data on the number of students identified by teachers as struggling writers, but not scoring at or below the 25th percentile, were not collected.
2. The G coefficients for relative decisions were virtually identical to those for G coefficients for absolute decisions (they never differed by more than one hundredth of a point). Thus, the same interpretations that can be drawn for absolute decision Table 4 can also be drawn for relative decisions.

References

Baker, E., Abedi, J., Linn, R., & Niemi, D. (1996). Dimensionality and generalizability of domain-independent performance assessments. *Journal of Educational Research*, *89*, 197–205.

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, *74*, 29–58.

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

Brown, G., Glasswell, K., & Harland, D. (2004). Accuracy in scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*, 105–121.

Calfee, R. C., & Wilson, K. M. (2004). A classroom-based writing assessment framework. In C. A. Stone, E. R. Silliman, B. J. Ehren, & K. Apel (Eds.), *Handbook of language and literacy: Development and disorders* (pp. 583–599). New York, NY: The Guilford Press.

Calkins, L. (1986). *The art of teaching writing*. Portsmouth, NH: Heinemann.

Coffman, W. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement*, *3*, 151–156.

Common Core State Standards: National Governors Association and Council of Chief School Officers. (2010). Available from http://www.corestandards.org/

Cronbach, L., Gleser, G., Nanda, H., & Rajaratnam, N. (1972).
The dependability of behavioral measurements: Theory of
generalizability of scores and profiles. New York, NY: Wiley. Diederich, P. (1966). How to measure growth in writing ability.
English Journal, *55*, 435–449. Donovan, C., & Smolkin, L. (2006). Children's understand-
ing of genre and writing development. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 131–143). New York, NY: Guilford.

Durst, R., & Newell, G. E. (1989). The uses of functions: James Britton's category system and research on writing. *Review of Educational Research*, *58*, 375–394.

Englert, S., Raphael, T., Fear, K., & Anderson, L. (1988). Students' metacognitive knowledge about how to write informational texts. *Learning Disability Quarterly*, *11*, 18–46.

Field, A. (2000). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*. Thousand Oaks, CA: SAGE.

Gearhart, M., Herman, J. L., Novak, J. R., & Wolf, S. A. (1995). Toward the instructional utility of large-scale writing assess- ment: Validation of a new narrative rubric. *Assessing Writing*, *2*, 207–242.

Gillespie, A., Olinghouse, N., & Graham, S. (2013). Fifth grade students' knowledge about writing process and writing genres. *Elementary School Journal*, *113*, 565–588.

Graham, S. (2006). Writing. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (pp. 457–478). Mahwah, NJ: Erlbaum.

Graham, S., Harris, K. R., & Fink-Chorzempa, B. (2002). Contributions of spelling instruction to the spelling, writ- ing, and reading of poor spellers. *Journal of Educational Psychology*, *94*, 669–686.

Graham, S., Harris, K. R., Fink-Chorzempa, B., & MacArthur, C. (2003). Primary grade teachers' instructional adaptations for struggling writers: A national survey. *Journal of Educational Psychology*, *95*, 279–292.

Graham, S., Harris, K. R., & Hebert, M. (2011a). *Informing writ- ing: The benefits of formative assessment*. Washington, DC: Alliance for Excellence in Education.

Graham, S., Harris, K. R., & Hebert, M. (2011b). It is more than just the message: Analysis of presentation effects in scoring writing. *Focus on Exceptional Children*, *44*, 1–12.

Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and motivation of strug- gling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology*, *30*, 207–241.

Graham, S., & Hebert, M. (2011). Writing-to-read: A meta-analy- sis of the impact of writing and writing instruction on reading. *Harvard Educational Review*, *81*, 710–744.

Graham, S., Hebert, M., & Harris, K. R. (2011). Throw em' out or make em' better? High-stakes writ-

ing assessments. *Focus on Exceptional Children*, *44*, 1–12.
Graves, D. (1983). *Writing: Teachers and children at work*. Exeter, NH: Heinemann.
Hammill, L. A., & Larsen, S. (Eds.). (1996). *Test of written lan- guage–3 (TOWL-3)*. Austin, TX: ProEd.
Harris, K. R., Graham, S., & Mason, L. (2006). Improving the writing, knowledge, and motivation of struggling young writers: Effects of self-regulated strategy development with and without peer support. *American Educational Research*
Journal, *43*, 295–340. Harris, K. R., Lane, K., Driscoll, S., Graham, S., Wilson, K.,
Sandmel, K., . . . Schatschneider, C. (2012). Teacher- implemented class-wide writing intervention using self-reg- ulated strategy development for students with and without behavior concerns. *Elementary School Journal*, *113*, 160–191.
Hayes, J. (2011). Kinds of knowledge-telling: Modeling early writing development. *Journal of Writing Research*, *3*, 73–92. Hess, C. W., Haug, H. T., & Landry, R. G. (1989). The reliability of type-token ratios for the oral language of school age chil- dren. *Journal of Speech and Hearing Research*, *32*, 536–540. Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? A generalizability theory
approach. *Assessing Writing*, *13*, 201–218. Jeffery, J. V. (2009). Constructs of writing proficiency in US
state and national writing assessments: Exploring variability.
Assessing Writing, *14*, 3–24. Klein, P. D., & Rose, M. A. (2010). Teaching argument and expla-
nation to prepare junior students for writing to learn. *Reading*
Research Quarterly, *45*, 433–461. Lin, S. C., Monroe, B. W., & Troia, G. (2007). Development of
writing knowledge in Grades 2-8: A comparison of typically developing writers and their struggling peers. *Reading & Writing Quarterly*, *23*, 207–230.
Loomis, S. C., & Bourque, M. L. (2001). *National Assessment of Educational Progress achievement levels 1992-1998 for writing*. Retrieved from http://nces.ed.gov/nationsreportcard/ writing/achieve. asp
Lyon, G. R. (1996). Learning disabilities. *The Future of Children*, *6*, 54–76.
National Center for Education Statistics. (2012). *The nation's report card: Writing 2011*. Washington, DC: Institute of Educational Sciences, U.S. Department of Education.
National Commission on Writing. (2004). *Writing: A ticket to work . . . or a ticket out*. Retrieved from http://www.writing- commission.org/prod_downloads/writingcom/writing-ticket- to-work. pdf
National Commission on Writing. (2005). *Writing: A powerful message from state government*. Retrieved from http://www. writingcommission.org/prod_downloads/writingcom/power- ful-message-from-state.pdf
Novak, J., Herman, J., & Gearhart, M. (1996). Establishing the validity for performance-based assessments: An illustration for collections of students' writing. *Journal of Educational Research*, *89*, 220–233.
Nunnally, J. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.
Saddler, B., & Graham, S. (2007). The relationship between writing knowledge and writing performance among more and less skilled writers. *Reading & Writing Quarterly*, *23*, 231–247.
Schoonen, R. (2005). Generalizability of writing scores: An appli- cation of structural equation modeling. *Language Testing*, *22*, 1–30.
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE.
Slavin, R., Madden, N., & Karweit, N. (1989). Effective programs for students at risk: Conclusions for practice and policy. In R. Slavin, N. Karweit, & N. Madden (Eds.), *Effective programs for students at risk* (pp. 21–54). Boston, MA: Allyn & Bacon.
Swartz, C., Hooper, S., Montgomery, J., Wakely, M., Kruif, R., Reed, M., . . . White, K. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytic scoring methods. *Educational and Psychological Measurement*, *59*, 492–506.
Yuan, Y. C. (1990). *Multiple imputation for missing data: Concepts and new*