Read Writ (2017) 30:1287–1310
DOI 10.1007/s11145-017-9724-6

Writing evaluation: rater and task effects on the reliability of writing scores for children in Grades 3 and 4

Young-Suk Grace Kim[1] · Christopher Schatschneider[2] · Jeanne Wanzek[3] · Brandy Gatlin[4] · Stephanie Al Otaiba[5]

Published online: 6 February 2017

Abstract We examined how raters and tasks influence measurement error in writing evaluation and how many raters and tasks are needed to reach a desirable level of .90 and .80 reliabilities for children in Grades 3 and 4. A total of 211 children (102 boys) were administered three tasks in narrative and expository genres, respectively, and their written compositions were evaluated in widely used evaluation methods for developing writers: holistic scoring, productivity, and curriculum-based writing scores. Results showed that 54 and 52% of variance in narrative and expository compositions were attributable to true individual differences in writing. Students' scores varied largely by tasks (30.44 and 28.61% of variance), but not by raters. To reach the reliability of .90, multiple tasks and raters were needed, and for the reliability of .80, a single rater and multiple tasks were needed. These findings offer important implications about reliably evaluating children's writing skills, given that writing is typically evaluated by a single task and a single rater in classrooms and even in some state accountability systems.

Keywords Generalizability theory · Task effect · Rater effect · Assessment · Writing

✉ Young-Suk Grace Kim
youngsk7@uci.edu

[1] University of California, Irvine, 3500 Education Building, Irvine, CA 92697, USA

[2] Florida Center for Reading Research, Florida State University, Tallahassee, FL, USA

[3] Vanderbilt University, Nashville, TN, USA

[4] Georgia State University, Atlanta, GA, USA

[5] Southern Methodist University, Dallas, TX, USA

Introduction

Writing is a critical skill for success in academic achievement and in most careers (Graham, Harris, & Hebert, 2011). Thus, it is troubling that the majority of children (72%) in Grade 4 in the United States write at basic or below basic level and only 28% of students write at a proficient level in the most recent National Assessment of Educational Progress (NAEP) writing assessment (National Center for Education Statistics [NCES], 2003). It is therefore not surprising that the rigor of writing standards has received much attention at the elementary level in the Common Core State Standards (CCSS, National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) and other similar state standards in the United States.

Writing is included in the high stakes state accountability systems in the majority of states in the United States. In many states, fourth grade is the first occasion in which students participate in these tests, and writing typically receives greater instructional attention in Grades 3 and 4 (Beck & Jeffery, 2007; Graham et al., 2011). Despite increased attention on improving writing and on high stakes writing assessment, there is relatively limited research on writing evaluation for children in elementary school as the vast majority of previous studies about writing evaluation have been conducted with older or college-age students (e.g., Bouwer, Beguin, Sanders, & van den Bergh, 2015; Brennan, Gao, & Colton, 1995; Hale, Taylor, Bridgeman, Carson, Kroll, & Kantor, 1996; Gebril, 2009; Hamp-Lyons, 2007; Moore & Morton, 1999; Schoonen, 2005; Swartz et al., 1999; van den Bergh, De Maeyer, van Weijen, & Tillema, 2012; Weigle, 1998).

Writing evaluation

The ultimate goal of writing assessment is accurately evaluating students' writing proficiency. However, for a complex skill like writing, in addition to students' writing proficiency itself, multiple factors such as raters, tasks, and prompts influence students' writing scores (see Schoonen, 2012 for a discussion on writing assessment with regard to validity and generalizability). Raters vary in their interpretation of rubrics despite training; and tasks and prompts vary in the demands (e.g., Gebril, 2009; Schoonen, 2005; Swartz et al., 1999; see below for details). Minimizing these variations that are not relevant to the ultimate construct (i.e., construct irrelevant variance or measurement error) is key to precise evaluation of writing in order to generalize writing scores into the proficiency we infer—writing proficiency (Schoonen, 2012). To this end, it is imperative to have an accurate understanding of the amount of variation attributable to various sources in writing scores such as true individual differences, and differences due to raters and tasks. Thus, the goal of the present study was to expand our understanding about writing evaluation by examining the extent to which various factors such as raters and tasks influence the reliability of writing scores for widely used writing evaluation methods for elementary grade students, and examining the optimal number of raters and tasks needed for consistent results in writing scores.

Writing proficiency is defined and evaluated in multiple ways for different purposes. Consequently, various approaches have been used to evaluate developing writers' proficiency, including holistic scoring, productivity, and curriculum-based measurement (CBM) writing, and these have been included in the present study. Holistic scoring is widely used in research for developing writers (Espin, De La Paz, Scierka, & Roelofs, 2005; Olinghouse, 2008) as well as in national and state assessment including NAEP writing and high-stakes state writing tests. In holistic scoring, a single score is assigned to a child's written composition after considering multiple aspects such as quality of ideas, organization, spelling and writing conventions. Writing productivity (also called fluency in some studies) is the amount of writing, and has been frequently examined in studies with elementary grade children. Although amount of writing itself is not the end goal of writing, productivity is an important aspect particularly for developing writers because children in elementary grades are still developing language and literacy skills (e.g., transcription) that constrain their writing skills, and a certain amount of writing is required to achieve quality. Writing productivity has been consistently shown to be associated with writing quality (Abbott & Berninger, 1993; Graham et al., 1997; Kim, Al Otaiba, Wanzek, & Gatlin, 2015; Olinghouse, 2008), and writing productivity indicators typically include the total number of words, ideas, and sentences (Abbott & Berninger, 1993; Kim et al., 2011, 2015; Kim, Al Otaiba, Sidler, Greulich, & Puranik, 2014; Puranik, Lombardino, & Altmann, 2007; Wagner et al., 2011).

Third, the CBM (curriculum-based measurement) writing scoring procedures have also been used as a means of screening and progress monitoring for developing writers including students in elementary and middle schools (Coker & Ritchey, 2010; Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002; Gansle et al., 2004; Jewell & Malacki, 2005; Lembke, Deno, & Hall, 2003; McMaster & Campbell, 2008; McMaster, Du, & Pestursdottir, 2009; McMaster et al., 2011). CBM writing scores are purported to provide global indicators of students' writing performance for developing writers in order to identify students who need further attention in assessments and instruction (Deno, 1985). CBM writing scores include number of words written, correct word sequences ("any two adjacent, correctly spelled words that are acceptable within the context of the sample," McMaster & Espin, 2007, p. 70), incorrect word sequences, and incorrect words (see below for details; Graham et al., 2011; McMaster & Espin, 2007). These various scores capture different aspects. Correct and incorrect word sequences capture not only transcription skills and knowledge of writing conventions (i.e., capitalization and spelling) but also oral language skills such as grammatical accuracy (i.e., students' ability to generate words that are meaningful and grammatically correct in context). Incorrect words reflect students' ability in spelling and punctuation. Studies have demonstrated reliability and validity for CBM writing scores (Coker & Ritchey, 2010; Gansle et al., 2002, Jewell & Malacki, 2005; Kim et al., 2015; Lembke et al., 2003; McMaster et al., 2009, 2011). A recent study has shown that CBM writing, using derived scores such as correct minus incorrect word sequences, is closely associated with writing quality but a dissociable construct (Kim et al., 2015).

Reliability of scores in writing evaluation

Establishing reliability is particularly challenging in writing evaluation as multiple factors are likely to influence scores (Bouwer et al., 2015). In holistic scoring, students' written composition is evaluated based on an a priori established rubric, yet even very carefully prepared rubrics are open to some interpretation. For instance, one aspect to consider in the NAEP writing evaluation is the extent of story or idea development (NCES, 1999). Out of the possible scale of 1–6, a score of 5 is described as "tell(ing) a clear story with some development, including some relevant descriptive details." A score of 4, on the other hand, "tells a story with little development; has few details." Then, raters have to determine what a "clear" story is, and what "some" versus "little" development means in order to differentiate a score of 4 from 5. In addition, because multiple aspects are considered in holistic scoring, raters might vary in extent to which different aspects (e.g., content and idea development, vs. spelling and writing conventions) are deemed to be important in determining the score. Therefore, differences among raters, even with training, are likely to influence the student's score to some extent, and consequently, students' writing scores would vary as a function of who rates their writing. Indeed, studies have consistently shown that raters vary in terms of leniency or rigor of applying a scoring rubric and their views on importance of various aspects (Cumming, Kantor, & Powers, 2002; Eckes, 2008; Kondo-Brown, 2002), and this variation among raters contributes to inconsistency in writing scores (i.e., measurement error).

The rater effect appears to vary with the specific traits being evaluated (e.g., content and organization vs. spelling) and scoring procedures (holistic and counting number of words). For instance, in Lane and Saber's (1989) study, eight raters scored written compositions of 15 students in Grades 3–8 on four dimensions: ideas, development and organization, sentence structure, and mechanics. Scores were on a scale from 1 to 7. Their results revealed that approximately 12% of the variance in writing was attributable to person by rater interaction. Approximately 6% of variance in writing scores was attributed to various traits or dimensions such as ideas and mechanics (Lane & Sabers, 1989). In another study with developing writers, Swartz et al. (1999) examined 20 written samples from middle school students on the following dimensions of writing used in the Test of Written Language-2nd Edition (TOWL-2): thematic maturity, contextual vocabulary, syntactic maturity, contextual spelling, and contextual style. They found a large rater effect (33% of variance) in thematic maturity (the number of ideas represented in writing and a small rater effect (3% of variance) in vocabulary use (the number of words with seven or more letters). Although in both dimensions—thematic maturity and vocabulary use—the rater was asked to count, which might appear to be less vulnerable to measurement error than rating, there was a large difference in terms of inconsistency in scoring. The large rater effect in thematic maturity is concerning because thematic maturity is often considered an important aspect of writing quality (Hammill & Larsen, 1996; Kim et al., 2015; Olinghouse & Graham, 2009; Wagner et al., 2011). Moreover, when teachers evaluated elementary grade children's writing on various dimensions, even after 3 h of training, inter-rater reliability was low with only about 53–59% of writing samples receiving the same score by

different raters. In contrast, CBM writing scores, which was rated by graduate students (amount of training unspecified), had higher inter-rater reliability, ranging from approximately 82–98% (Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006).

The picture becomes more complex as writing scores are influenced by different tasks (Cooper, 1984; Gebril, 2009; Huot, 1990; Lane & Sabers, 1989; Schoonen, 2005; Swartz et al., 1999). A substantial amount of variation in writers' scores has been attributed to the interaction between person (writer) and task such that individuals' writing scores vary largely as a function of tasks. Approximately 21–53% of variance in various writing scores has been attributed to person by task interaction for children in Grade 6 (Schoonen, 2005) and approximately 21 and 22% for college students writing in second or foreign language (L2; Gebril, 2009). The task effect may be due to the writer's background knowledge in relation to the task. These results indicate that using a single task would not yield reliable information about students' writing skill (Graham et al., 2011).

Finally, narrative and expository genres differ in terms of children's skills, experiences, and familiarity (Bouwer et al., 2015; Olinghouse, Santangelo, & Wilson, 2012). Many elementary grade children are more familiar with narrative texts than informational/expository texts (Duke, 2014; Duke & Roberts, 2010).Children's performance on writing varies as a function of genre and therefore, children's performance on one genre cannot necessarily be generalized to another (see Graham et al., 2011; Olinghouse et al., 2012). Instructionally, in the US where the present study was conducted, narrative and expository genres are typically taught somewhat separately. Narrative genres are typically introduced earlier and taught more frequently in primary grades. Standards in writing, such as the widely used Common Core State Standards, also specify goals for each genre. Therefore, although both narrative and expository writing are part of an overall writing skill, it is important to examine whether factors influencing students' writing scores vary as a function of genre.

Generalizability theory

The present study used generalizability theory (GT) to address the primary research question about amount of variance attributable to different factors or facets such as raters and tasks. GT partitions variance into multiple sources of error variance (called facets), and interactions among these sources simultaneously as well as true variance among individuals (Shavelson & Webb, 1991, 2005). In contrast, in the classical test theory, variance of observed scores are partitioned to two estimates—true score variance and error variance—and only one type of error variance is captured at a time (e.g., rater; Swartz et al., 1999). Another important aspect of GT is that it allows examination of reliability of decision studies (also called Decision or D studies) to use the variance components in the GT to inform measurement features that can help minimize the measurement error (Bachman, 2004; Cronbach, Gleser, Nanda, & Rajaratnum, 1972). Partitioned variance can be used to estimate how to minimize the effect of error variance for different purposes such as making relative or absolute decisions about students (i.e., rank ordering students based on

writing performance or deciding whether students' writing meets a particular criterion level of performance). For instance, decision study results inform researchers and educators about how many raters or tasks are necessary to reach a certain level of reliability for either relative or absolute decisions.

Despite accumulating evidence about factors related to the reliability of writing scores, overall there is particularly limited research base about factors influencing the reliability of writing scores and the extent of their influences for beginning writers. Many previous studies were conducted with college students or adult learners in English as a second or foreign language contexts (e.g., Barkaoui, 2007; Cumming et al., 2002; East, 2009; Gebril, 2009; Weigle, 1998) or in languages other than English (e.g., Bouwer et al., 2015; Eckes, 2008; Kondo-Brown, 2002; Kuiken & Vedder, 2014; Schoonen, 2005). The few existing studies with beginning writers in English had small sample sizes (e.g., 20 writing samples in Swartz et al., 1999; 15 children in Lane & Sabers, 1989). Importantly, although holistic, productivity, and CBM writing scores are widely used for various purposes and contexts for developing writers, they have not been examined for score reliability for elementary grade students. This is an important gap in the literature, given the recognized importance of writing in elementary grades (e.g., Common Core State Standards), and inclusion of writing in the high stakes assessment in the elementary grades.

Understanding factors influencing the reliability of writing scores for children in elementary grades has an important implication in various contexts. In a high stakes context (e.g., state level writing proficiency test), unreliable assessment would have an important consequence by incorrectly identifying children who meet or do not meet required proficiency levels. The reliability of writing scores is also important in the instructional or classroom context, a relatively low stakes context. Given high stakes writing tests in Grade 4 in many states as well as explicit elaboration of benchmark on writing skills in the Common Core State Standards or similar standards, writing has received increased attention in instruction particularly in Grades 3 and 4. Thus, teacher decisions on students' writing score have implications because those scores are used to determine who is eligible for additional or more intensive instruction as well as determining students' disability status in writing.

Our goal in the present study was to examine the extent to which multiple factors such as raters and tasks influence writing scores in narrative and expository genres with widely used evaluation approaches (i.e., holistic, productivity, and CBM writing) and how many raters and tasks are needed to reliably evaluate writing skill in these approaches for children in Grades 3 and 4. This question was addressed using a generalizability theory approach which allows disentangling multiple sources of error, and interactions among these sources. Among several sources of variation, the raters and tasks were examined based on findings from previous studies.

Our study adds uniquely to the literature in several ways. First, we examined multiple approaches to writing evaluation that are widely used in research and high-stakes testing for children in elementary grades, including holistic scoring, productivity indicators, and curriculum-based measurement (CBM) writing scores. These various approaches tend to be used for different purposes (i.e., holistic scoring in high-stakes context and research; productivity in research, and CBM in

schools and research). Therefore, an understanding about the extent to which multiple sources of measurement error are manifested in these widely used writing evaluation approaches would be informative for writing evaluation for various stakeholders, including teachers, researchers, and policy makers. Although productivity and CBM writing scores might appear to be straightforward as counting certain targets, in fact, there are some aspects that raters have to consider (e.g., determining grammaticality; see below for details), and therefore, the rater aspect might contribute some variance in these scoring approaches. In addition, the extent to which the task facet contributes to the variance in the productivity and CBM writing scores would be informative, given its wide use in research and school setting. Second, we examined the effects of tasks and raters in both narrative and expository genres because students' performances on different genres are only moderately correlated (Graham et al., 2011; Olinghouse et al., 2012).

Finally, if multiple factors such as raters and tasks do influence students' writing scores, then it is crucial to determine how many raters and tasks are needed to reduce inconsistency (i.e., measurement error) to reach an acceptable level of reliability. Information about the optimal number of raters and tasks for acceptable or desirable reliabilities for different purposes is important for informing practices and for resource allocations (Schoonen, 2005). Therefore, we examined the effect of increasing the number of raters and tasks on measurement error for relative and absolute decisions. In relative decisions, the focus is on rank ordering persons according to performance levels as in normed and standardized writing tasks whereas in absolute decisions the absolute level of performance is the primary focus [i.e., meeting a prespecified target proficiency level as in 'criterion-referenced' assessment and classifying students into specified groups (Swartz et al., 1999)]. In the present study, .90 and .80 were set as the acceptable criterion reliabilities to examine how many raters and tasks are needed. The high .90 criterion was based on Nunnally (1967) and DeVellis (1991), and was based on the fact that the consequence of absolute decisions are critical and severe in a high-stakes state assessment context (e.g., holistic scoring used in the high stakes contexts). Furthermore, an alternative criterion reliability of .80 was examined because this is more practically feasible in many settings including research and classrooms.

Present study

The overall goal of the present study was to examine the effect of raters and writing tasks on the reliability of writing scores in widely used scoring procedures in various contexts (i.e., holistic scoring used in high-stakes context, productivity, and CBM scoring in classrooms) for children in Grades 3 and 4. The following were specific research questions:

1. What percentage of total score variance in holistic, productivity, and CBM writing scores is attributed to persons (i.e., students), raters, and tasks? How do the percentages vary across writing evaluation methods (holistic, productivity, and CBM writing)? How do the percentages vary in narrative and expository genres?

2. What is the effect of increasing the number of raters and tasks on score reliability for relative and absolute decisions? How many raters and tasks are needed to reach the reliabilities of .90 or .80?

We hypothesized that most of variance would be attributable to child's ability. However, rater and task effects were expected given consistent findings of these factors in prior studies with older students. We did not have a hypothesis about the specific number of raters and tasks needed for reliabilities of .90 and .80 other than the hypothesis that the number would vary depending on the evaluation approach. It should be noted that the goal of the present study was not to examine reliability or validity of the assessments used in the study (e.g., TOWL-4). Instead, we aimed to investigate the extent to which raters and tasks contribute to accuracy of writing scores when using various scoring approaches that are widely used in high stakes and low stakes contexts for children in elementary schools (i.e., holistic scoring, productivity, and CBM writing). Also note that these three evaluation methods do not necessarily represent constructs. For instance, total number of words is widely used both as productivity and CBM writing. In terms of dimensionality, total number of words was best described as a productivity measure—a related but dissociable construct from other derived CBM scores (e.g., correct word sequences minus incorrect word sequences; see Kim et al., 2015). However, in the field of writing research and practice in the classrooms, total number of words is typically included as part of CBM writing scores. Although the construct or dimensionality question is important (see Kim et al., 2014, 2015; Puranik, Lombardino, & Altmann, 2008; Wagner et al., 2011), it was beyond the scope of the present study. Instead, the goal of the present study was to evaluate reliability of various writing evaluation methods as they are used in research and practice.

Method

Participants

Data were collected from 211 children (102 boys) in Grades 3 ($n = 86$) and 4 ($n = 125$). These children were drawn from 68 classrooms in 18 schools in a mid-sized city in the southeastern part of the United States. These children were part of a larger longitudinal study of children's literacy development (see Kim et al., 2014). The participating schools varied largely in terms of socio-economic status of children they served. Mean ages were 8.23 ($SD = .36$) and 9.22 ($SD = .32$) for children in Grades 3 and 4, respectively. Approximately 42% of the children were Caucasians, and 43% African Americans, and the rest were multiracial or other racial minority (e.g., Asian). Approximately 68% of the children were eligible for free and reduced lunch, a proxy for low socioeconomic status, and 10% of the children were receiving special education services, most under the label of learning disabilities or language impairment. The schools used a district developed writers' workshop approach for their writing curriculum, which included the process of prewriting, drafting, teacher-student conference, revising, editing, and publication.

Instrument

Three tasks were used for the narrative and expository genres, respectively, with a total of six tasks. The narrative tasks included the Test of Written Language-4th edition (TOWL-4; Hammill & Larsen, 2009) as well as two experimental tasks (Magic Castle and One Day). For the story composition subtest of the TOWL-4, students heard or were read a story accompanying a full color picture read aloud by the assessor. Then, students were presented with another picture and instructed to write a story that goes with the picture. The Magic Castle task was adapted from the 1998 NAEP narrative task for Grade 4. In this task, the students were provided with the beginning of a story about a child who discovers a castle that has appeared overnight. The students were then told to write a story about who the child meets and what happens inside the castle. The One Day task ("One day when I got home from school…") has been used in previous studies (Kim et al., 2013, 2014; McMaster et al., 2009, 2011) and required the student to write a story about something unusual or interesting that happened to them. Previous studies have reported reliabilities using the experimental tasks ranging from .82 to .99 (Kim et al., 2015). For the TOWL-4, test–retest reliability was reported to be .70 (Hammill & Larsen, 2009).

Three expository tasks included the essay composition subtask of the Wechsler Individual Achievement Test-3rd edition (WIAT-3; Wechsler, 2009) as well as two experimental tasks (Librarian and Pet). In the WIAT-3 essay composition task, students were asked to write about their favorite game and include three reasons why they like it. The Librarian and Pet tasks were adapted from the NAEP Grade 4 tasks. For the Librarian essay task, students were told to imagine that their favorite book is missing from the library. Their task was to write a letter to the school librarian asking her to buy the book again. For the Pet task, students were instructed to write a letter to their parents explaining what animal they would like to have as a pet and why that animal would make a good pet (Wagner et al., 2011). Reliability using these tasks has been reported to range from .82 to .89 (Kim et al., 2015; Wagner et al., 2011). For the WIAT-3 essay composition task (i.e., game task), test–retest reliabilities were reported to range from .86 to .87 (Wechsler, 2009).

Procedures

Data collection

Data were collected in the fall (September and October) by trained research assistants. Children were assessed in groups of 6–8 students in three sessions. Following standard procedures, children had 15 min[1] to write in each writing task and all were administered with paper and pencil. Children were given two writing

[1] Children were given 15 min based on our experiences with elementary grade children. CBM writing assessments (e.g., writing tasks) typically have shorter assessment times (e.g., 3 min). This does not present a validity issue in the present study because the purpose of our study was examining reliability of various evaluation approaches including CBM writing indicators, not a particular CBM writing test (e.g., picture task) per se.

tasks at a time, one narrative and one expository each week for a total of 3 weeks. TOWL-4 writing (narrative) and WIAT-3 writing (expository) were administered in Week 1; Pet (expository) and One Day (narrative) tasks were administered in Week 2; and Magic Castle (narrative) and Librarian (expository) tasks were administered in Week 3.

Scoring procedure

Three different types of evaluation were conducted: holistic scoring, productivity indicators, and CBM writing scoring. In the holistic scoring, raters assigned a single score on a scale of 0–6 while taking into account several aspects such as ideas, organization, language use, and writing conventions (e.g., spelling and punctuation). Scoring guidelines were adapted from the publicly available Florida Comprehensive Assessment Test (FCAT) and NAEP scoring guidelines for Grade 4. For a score of 6, the student's composition had fully developed ideas with sufficient supporting details and clear organization, and appropriate and skilled language used with few spelling and punctuation errors. A score of 0 was assigned when the composition was simply a rewording of the task or the response was not related to the task at all, or the composition was illegible. Although the scoring guide had a score of 0–6, none of the writing samples received a score of 6 (see Table 2) and a score of 0 was rare (fewer than 6 students).

In order to capture productivity in writing, the number of sentences and ideas were counted by raters (see Kim et al., 2011, 2013, 2014; Puranik et al., 2008; Wagner et al., 2011). For the number of sentences, if periods were missing in students' compositions but contextually and linguistically a complete sentence, it was counted as a sentence. The number of ideas was a total number of propositions, which were defined as predicate and argument. For example, "I ate breakfast and went to school" was counted as two ideas.

The CBM scoring included the number of words written, correct word sequences, incorrect word sequences, and incorrect words (Coker & Ritchey, 2010; Lembke et al., 2003; McMaster et al., 2009, 2011; see Graham et al., 2011; McMaster & Espin, 2007, for reviews). The number of words written is a total number of words in the composition. Note that although number of words written is often part of CBM writing scores, it is not unique to the CBM scoring and has been widely used as a writing productivity indicator (Abbott & Berninger, 1993; Kim et al., 2011, 2014; Puranik et al., 2007; Wagner et al., 2011). In the present study, we report it under the CBM writing scores. Correct word sequences refers to two adjacent words that are grammatically correct and spelled correctly and the incorrect word sequences refers to any two adjacent words that are incorrect (McMaster & Espin, 2007). Incorrect words were any words that were spelled incorrectly resulting in a nonword (e.g., *favrit* for *favorite*) and words in which the first letter should have been capitalized, but was not.

Raters used a copy of students' original handwritten compositions without corrected spelling and punctuations and without student identification information. The use of original handwritten composition was important, particularly for CBM

writing scores because punctuations such as capitalization are taken into consideration for scoring.

Rater training procedure

Due to practical reasons of coding a large number of writing compositions (approximately 1200 compositions) in a reasonable time, raters were nested in some scoring types. In other words, different pairs of raters conducted different types of scoring (i.e., two raters in each type of scoring such as holistic scoring versus productivity and CBM writing). Note, however, the lead rater in productivity scoring and CBM scoring were the same. Therefore, a total of five raters were involved in three types of scoring (i.e., Raters 1 and 2 for holistic scoring; Raters 3 and 4 for productivity; and Raters 3 and 5 for CBM writing). All the raters were females. The holistic scoring was conducted by two raters, a graduate student in special education and an individual with an undergraduate degree in French. Both raters had experience working with children in terms of administering assessments and teaching children in a literacy intervention. The first author trained the two raters in an initial 3-h meeting in which the scoring manual was reviewed and discussed, and some sample compositions on one task were scored together. Children's writing samples from previous studies as well as publicly available scored writing anchor samples for FCAT and NAEP were utilized in the initial training. The two raters then independently rated 10 writing samples on a task, and reconvened with the first author to share the scores, discuss and refine the manual, and resolve differences in scores. This subsequent meeting was conducted separately for the narrative and expository tasks. The unique meeting for each task was done to ensure the scoring guidelines were consistently applied to all the six writing tasks. Total time spent on discussion, excluding raters' independent scoring of practice samples, across all the tasks for holistic scoring was 14 h (3 h initial meeting which included examples of 1 narrative task; 1 h meeting for each task to discuss independent practice samples; 1 h meeting for each task about applying the rubric consistently for each of five tasks $= 3 + 6 + 5$).

For the productivity scoring, two graduate students (one in school psychology and the other special education) conducted scoring. The student in school psychology had extensive experience with student assessment and scoring, including writing. The first author had an initial 2-h training to describe and discuss the scoring manual and procedures, and practice scoring. Then, the raters independently scored 10 children's writing samples from a previous study and met with the first author in a subsequent meeting to clarify scoring manual, and discuss scores and discrepancies in scores. A total of 5 h were spent in training and discussion.

For the CBM scoring, two graduate students (the one in school psychology and the other in communication disorders) conducted scoring. The graduate student in school psychology had an extensive training and experience in CBM scoring in previous studies and therefore trained the other graduate student. The two raters met initially for approximately 2 h in which training and discussion of the scoring manual and procedures took place. Next, the two raters each scored 20 practice

pieces and later met to discuss discrepancies and resolve differences in scores. A total of 5 h were spent in training and discussion in CBM scoring. As noted above, given that different pairs of raters conducted different types of scoring, caution needs to be taken in directly comparing rater effects across different writing evaluation methods.

Note that in a study examining factors influencing reliability, there is no a priori reliability to be met before raters proceed to code writing samples because the goal is to investigate how much variance is attributable to raters, given a specific amount of training. In typical studies of writing, reliability (i.e., inter-rater reliability) is established before raters evaluate children's written compositions independently. However, this would not permit researchers to achieve the goal of the present study–examining reliability attributable to various sources given the amount of training–if a certain level of reliability is already established (see previous work on generalizability theory). Previous studies varied largely in terms of amount of training provided to raters, ranging from 3 h (Kondo-Brown, 2002) to 6 h (Swartz et al., 1999). Importantly, many studies did not report the amount of training (Gebril, 2009; Knoch, 2009; Lane & Sabers, 1989; Schoonen, 2005, 2012; Stuhlmann, Daniel, Delinger, Denny, & Powers, 1999; Tillema, van den Bergh, Rijlaarsdam, & Sanders, 2012; van den Bergh et al., 2012). In other studies (e.g., Lane & Sabers, 1989), raters were only provided with written scoring guidelines, anchor essays, and practice essays with expert rater's scores without a formal training.

Generally, the amount of training would vary for different evaluation methods to reflect varying nature of demand to conduct evaluation. For instance, holistic scoring typically is more open to variation in how raters interpret rubrics than productivity (e.g., counting the number of sentences), and therefore, typically requires greater amount of training. In the present study, the amount of training described above represents when raters expressed their clarity about how to evaluate students' writing in a given scoring approach.

Data analysis

The primary data analytic strategy was the generalizability theory (GT), using variance analytic techniques (Cronbach et al., 1972; Shavelson & Webb, 1991; Shavelson, Webb, & Rowley, 1989). The present study had a two facet[2] fully crossed design (task × rater) with the following equation:

$$\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(tr) + \sigma^2(ptr)$$

A GT macro developed by Mushquash and O'Connor (2006) for use in SPSS was used to estimate the proportion of variance in these scores attributable to true differences among persons, and error in measurements that may be due to raters, task, the interaction of rater and task, and random error (see the equation above). Missing data were handled by the use of full-information maximum likelihood when

[2] Facets are measurement features or sources of variation such as person, rater, and task.

estimating the variance components, and persons, rater, and tasks were treated as random factors in this model.

Results

Table 1 shows the descriptive statistics in different genres and tasks by raters and evaluation types. The following were the number of students' written compositions available per prompt: TOWL ($n = 209$), Castle ($n = 198$), One day ($n = 200$), WIAT ($n = 208$), Librarian ($n = 197$), and Pet ($n = 202$). Missing data were due to student absences on the day of assessment. Mean scores in the holistic rating ranged from 1.78 to 2.47 with sufficient variation around the means. These mean scores are somewhat low, given that the range was from 0 to 6. This might be attributed to two facts: (1) the holistic rating was adapted from state and national assessments for Grade 4 students while the sample consisted of children in Grades 3 and 4; and (2) many of the participating students (approximately 68%) were from low socioeconomic families. We adapted Grade 4 rubric for state and national level assessment because that at the time of the study, the state-level high stakes writing rubric for Grade 4 was used in the elementary schools. Furthermore, studies have consistently shown that students from low SES backgrounds have lower writing skills (e.g., Applebee & Langer, 2006; National Center for Education Statistics, 2012). Students, on average, wrote approximately 87–111 words, 8–11 sentences, and 12–17 ideas. Although there was variation across tasks within each genre, students, on average, tended to write more in the narrative tasks than in the expository tasks.

Research question 1: Variance of scores (holistic, productivity, and CBM writing) attributed to persons, raters, and tasks

Tables 2, 3, and 4 show variance components in holistic, productivity, and CBM writing scores in narrative and expository genres, respectively. As shown in Table 2, the Person explained largest amounts of variance in holistic scores in both narrative and expository genres. The amount of variance was similar in narrative and expository tasks (54 and 52% in narrative and expository tasks, respectively). These results indicate that true individual differences among children explain the largest amount of variability in narrative and expository writing scores when using holistic rating scores. The second largest component was the Person × Task interaction with similar amounts of variance, 30.44 and 28.61% of variance explained in narrative and expository tasks. These results indicate that rank ordering of children differed to a large extent by task. The residual variance explained the third largest amount with 10.78 and 12.62% in narrative and expository tasks, respectively. Other components such as Task, Rater, or Rater × Task explained only a small or minimal amount of variance.

Similar patterns were observed in the productivity (Table 3) and CBM scores (Table 4). For the productivity indicators, the largest amount of variance was attributable to Person (48–56%), followed by the interaction between Person and Task (43–48%). Similarly, the largest component in the CBM scoring was Person,

Table 1 Descriptive statistics in writing scores

Genre and prompt	Holistic scoring		Productivity			
			# of sentences		# of ideas	
	M (*SD*)	Min–max	*M* (*SD*)	Min–max	*M* (*SD*)	Min–max
Narrative						
TOWL	1.78 (.91)	0–5	10.66 (4.73)	0–28	15.94 (6.33)	0–39
Castle	2.18 (1.00)	0–5	11.45 (5.33)	1–29	16.94 (7.64)	1–40
One Day	2.17 (1.00)	0–5	10.92 (5.41)	0–32.5	16.56 (7.47)	1–39
Expository						
WIAT	2.43 (1.00)	0–5	9.16 (4.58)	0–26	14.17 (6.79)	0–39.5
Librarian	2.01 (.90)	0–5	7.73 (4.28)	0–24.5	12.30 (6.04)	1–33.5
Pet	2.47 (.99)	0–5	8.15 (4.06)	0–23.5	14.44 (6.73)	0–37

Genre and prompt	CBM writing							
	Total words written		Correct word sequences		Incorrect word sequences		Incorrect words	
	M (*SD*)	Min–max	*M* (*SD*)	Min–max	*M* (*SD*)	Min–max	*M* (*SD*)	Min–max
Narrative								
TOWL	105.16 (40.65)	0–236	82.16 (38.78)	0–203.5	33.85 (20.76)	0–117	11.02 (7.89)	0–49
Castle	111.42 (51.37)	3–282.5	86.39 (47.17)	1–228.5	36.86 (23.47)	1–128	11.77 (8.76)	0–50.5
One Day	104.81 (47.89)	8–245	84.71 (44.11)	4–235.5	31.06 (19.67)	0–106	9.35 (6.88)	0–48
Expository								
WIAT	99.44 (46.99)	18–271	81.15 (41.29)	7–202.5	29.01 (20.59)	0–125.5	9.80 (8.17)	0–58
Librarian	86.63 (42.54)	2.50–232	69.14 (38.04)	2–210	26.17 (18.53)	0–112.5	9.26 (7.41)	0–46
Pet	96.83 (47.07)	9–258	81.01 (43.38)	1–228	24.88 (18.82)	0–105.5	7.86 (6.81)	0–50

CBM curriculum-based measurement, # = number

Table 2 Estimated percent variance explained in holistic ratings of narrative and expository writing tasks

Variance component	Narrative	Expository
Person	54.24	51.81
Rater	0	0
Task	4	6
Person × Rater	0	.5
Person × Task	30.46	28.57
Rater × Task	.1	.4
Residual	10.78	12.62
G coefficient		
Relative (G)	.82	.81
Absolute (Phi)	.80	.79

Relative relative decision, *Absolute* absolute decision

Table 3 Estimated percent variance explained in productivity indicators of narrative and expository writing tasks

Variance component	Number of sentences		Number of ideas	
	Narrative	Expository	Narrative	Expository
Person	53	48	56	50
Rater	0	0	.1	0
Task	.2	2.5	.1	2.9
Person × Rater	0	0	0	0
Person × Task	45	48	43	46
Rater × Task	0	0	0	0
Residual	1.7	1.4	1.1	.8
G coefficient				
Relative (G)	.77	.75	.79	.76
Absolute (Phi)	.77	.74	.79	.75

Relative relative decision, *Absolute* absolute decision

explaining approximately 51–69% of total variance. The second largest component was the Person × Task interaction, explaining 31–45% of variance in children's writing score. Variability due to Rater, Person × Rater, or residual variance was very small or minimal in all the productivity and CBM writing scores (see Table 4).

Research question 2: Effect of number of raters and tasks on reliability

Tables 2, 3, and 4 also display generalizability and phi coefficients from a series of decision studies. The generalizability coefficient is relevant to relative decisions, and is the ratio of universe score variance to the universe score variance and the relative error variance (Brennan, 2011). The phi coefficient is relevant to absolute decisions, and is the ratio of universe score variance to the universe score variance

Table 4 Estimated percent variance explained in CBM indicators of narrative and expository writing tasks

Variance component	Number of words		Correct word sequences		Incorrect word sequences		Incorrect words	
	Narrative	Expository	Narrative	Expository	Narrative	Expository	Narrative	Expository
Person	61	57	69	63	59	56	57	51
Rater	0	0	0	0	.2	.3	0	0
Task	.3	2	0	3	1	.3	2	1.4
Person × Rater	0	0	0	.1	0	.4	0	.3
Person × Task	38	41	31	34	37	41	38	45
Rater × Task	0	0	0	0	0	0	0	0
Residual	0	0	.5	.5	2.2	2.3	3	2.5
G coefficient								
Relative (G)	.83	.81	.87	.85	.82	.80	.81	.77
Absolute (Phi)	.83	.80	.87	.84	.82	.80	.81	.76

Relative relative decision, *Absolute* absolute decision

and the absolute error variance. When interpreting these results, it should be kept in mind that generalizability coefficients reported in Tables 2, 3, and 4 are based on the current study design of 2 raters and 3 writing tasks (tasks) in each genre, and the described amount of training of raters.

In holistic scores, the generalizability coefficient was .82 in the narrative tasks, and .81 in the expository tasks. The phi coefficient was .80 and .79 in the narrative and expository tasks, respectively. The generalizability coefficients for the productivity indicators ranged from .75 to .79 whereas phi coefficients ranged from .74 to .79. The generalizability and phi coefficients for CBM writing scores ranged from .76 in the number of incorrect words of expository tasks to .87 in correct word sequences of narrative tasks. The finding that phi coefficients were lower than generalizability coefficients is in line with other studies (e.g., Gebril, 2009; Schoonen, 2005). Recall that generalizability coefficients are for relative decisions and phi coefficients are for absolute decisions. Therefore, relative decisions (i.e., rank-ordering children) are more relevant to standardized and normed tasks where the primary goal is to compare a student's performance to that of the norm sample. Absolute decisions are relevant to dichotomous, criterion-referenced decisions such as classifying children as proficient and not proficient, as in high-stakes testing or determining which students require supplementary writing instruction in the classroom contexts.

In order to examine the effect of increasing the number of tasks and raters on score reliability, decision studies were conducted. To reach the criterion reliability of .90, when using holistic scoring, a minimum of 2 raters and 6 tasks were needed for relative decisions, and 2 raters and 7 tasks or 4 raters and 6 tasks were needed for absolute decisions in the narrative genre. In the expository genre, a minimum of 2 raters and 6 tasks were needed for relative decisions, and 3 raters and 7 tasks were needed in the expository genre. For productivity scores, at least 1 rater and 7 tasks were needed for relative and absolute decisions in the narrative genre whereas greater than 7 raters and 7 tasks were needed in the expository genre. When using CBM scores, a minimum of 1 rater and 6 tasks are needed in both genres for the total number of words. For the correct word sequences, 1 rater and 4 tasks were needed for both relative and absolute decisions in the narrative genre whereas in the expository genre, a minimum of 1 rater and 5 tasks were needed for relative decisions, and 1 rater and 6 tasks were needed for absolute decisions. Somewhat similar patterns were observed for the incorrect word sequences and incorrect words.

To reach the criterion of .80 reliability, in holistic scoring, a single rater and 3–4 tasks were necessary, depending on the narrative versus expository, and types of decisions. In productivity scoring, 4 tasks were required with a single rater. Similar patterns were observed for different outcomes for CBM writing scores, ranging from 2 to 4 tasks with a single rater.

Figures 1 and 2 illustrate results of holistic scoring and the number of sentences outcome (productivity scoring), respectively. Results of CBM scores are not illustrated with a figure because of its highly similar pattern to Fig. 2. These figures illustrate a large effect of tasks and a minimal effect of raters on score

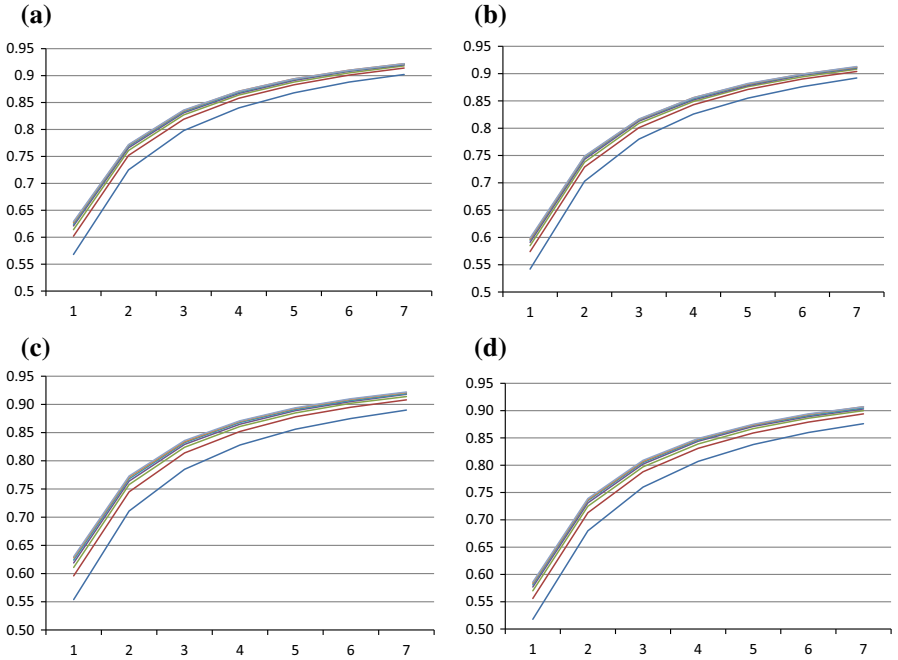

Fig. 1 Generalizability and phi coefficients of holistic scores as a function of raters and tasks: *Y* axis represents reliability; *X* axis represents number of tasks; *lines* represent number of raters from one rater (*lowest line*) to seven raters (*highest line*). **a** Generalizability coefficient in narrative genre. **b** Phi coefficient in narrative genre. **c** Generalizability coefficient in expository genre. **d** Phi coefficient in expository genre

reliability. It is clear that increasing the number of tasks (x axis) had a large return in score reliability.

Discussion

In this study, we examined the extent to which raters and tasks influence the reliability of various methods of writing evaluation (i.e., holistic, productivity, and CBM writing) in both narrative and expository genres, and the effect of increasing raters and tasks on reliability for relative and absolute decisions for children in Grades 3 and 4. For the latter question, criterion reliabilities were set at .90 and .80.

Overall, the largest amount of variance was attributable to true variance among individuals, explaining 48–69% of total variance. However, a large person by task effect was also found, suggesting that children's writing scores varied by tasks to a large extent, explaining 29–48% of variance. This was true across narrative and expository genres, and various evaluation methods including holistic scoring, productivity indicators such as number of sentences and number of ideas, and CBM writing scores such as correct word sequences, incorrect word sequences, and incorrect words. The large task effect was also evident in the decision studies, and

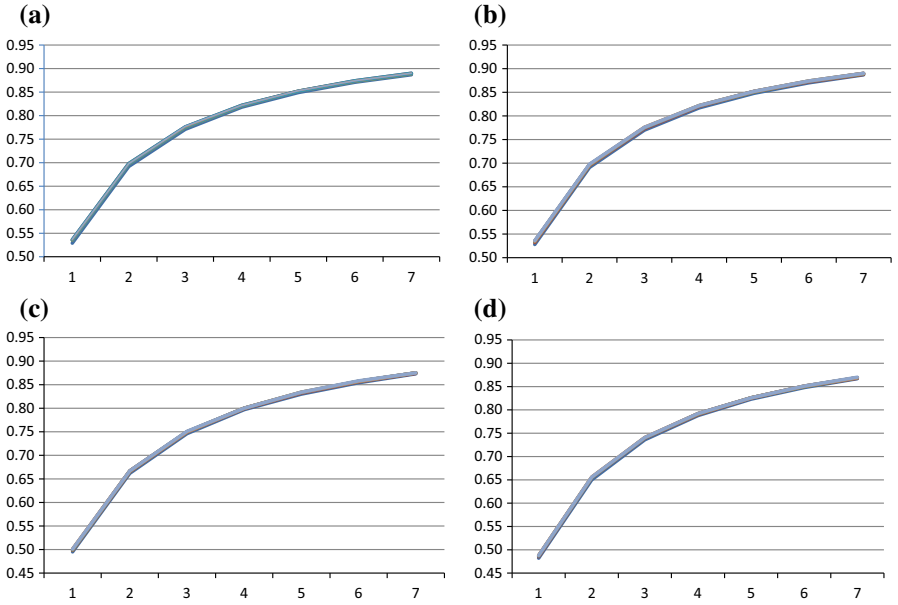

Fig. 2 Generalizability and phi coefficients of number sentences as a function of raters and tasks: *Y* axis represents reliability; *X* axis represents number of tasks; *lines* represent number of raters from one rater to seven raters (*lines* largely overlap due to small rater effect). **a** Generalizability coefficient of narrative genre. **b** Phi coefficient of narrative genre. **c** Generalizability coefficient of expository genre. **d** Phi coefficient of expository genre

increasing the number of tasks had a substantial effect on improving reliability estimates. To reach a desirable level of reliability of .90, a large number of tasks were needed for all the scoring types although some variation existed among evaluative methods. For instance, in holistic scoring, a minimum of 6 tasks and 4 raters and, or 7 tasks and 2 raters were needed for absolute decisions in the narrative genre. In addition, a minimum of 4–6 tasks was needed for correct word sequences of CBM scoring for both relative and absolute decisions. When the criterion reliability was .80, approximately 2–4 tasks were required with a single rater. The large task effect is line with relatively weak to moderate correlations in children's performance on various writing tasks (see Graham et al., 2011). One source of a large task effect is likely to be variation in background knowledge, which is needed to generate ideas on topics in the tasks (Bereiter & Scardamalia, 1987). The tasks used in the present study were from normed and standardized tasks as well as those used in previous research studies, and the tasks were not deemed to rely heavily on children's background knowledge. For instance, the narrative tasks (i.e., TOWL-4, Magic castle, One day) involved experiences that children are likely to have in daily interactions. Similarly, topic areas in the expository tasks were expected to be familiar to children such as favorite game, requesting a book to the librarian, and a pet. Nonetheless, children are likely to vary in the extent of richness in experiences related to these topic areas as well as the extent to which they can utilize this background knowledge in writing. The large task effect is consistent with a previous

study with older children in Grade 6 (e.g., Schoonen, 2005), and highlights the importance of including multiple tasks in writing assessment across different evaluation methods.

In contrast to the task effect, the rater effect was minimal in all the different evaluative methods. This minimal effect of rater is divergent from previous studies (e.g., Schoonen, 2005; Swartz et al., 1999). As noted above, previous studies have reported mixed findings about a rater effect, some reporting a relatively small effect whereas others report a large effect (e.g., 3–33%). We believe that one important difference between the present study and previous studies is the amount of training raters received, which consisted of an initial training, independent practice, followed by subsequent meetings. In particular, for holistic scoring, a subsequent meeting occurred for each task to ensure consistency of application of the rubric to different writing tasks. Overall, a total of 24 h were spent on training of holistic scoring, productivity, and CBM writing. As noted earlier, previous studies either reported a small amount of training (3–6 h on 4–5 dimensions) or did not report amount of training (e.g., Kondo-Brown, 2002; Swartz et al., 1999). The amount of training is an important factor to consider because training does increase the reliability of writing scores (Stuhlmann et al., 1999; Weigle, 1998). Therefore, the rater effect is likely to be larger when raters do not receive rigorous training on writing evaluation. A future study is needed to investigate the effect of rigor of training on reliability for different evaluation methods and to reveal the amount of training needed for evaluators of various backgrounds (e.g., teachers) to achieve adequate levels of reliability.

These findings, in conjunction with those from previous studies, offer important implications for writing assessments at various levels—state level high-stakes assessments as well as educators (e.g., teachers and school psychologists) working directly with children and involved in writing evaluation. It is not uncommon that a child's written composition is scored by a single rater, even in high-stakes testing. Although the rater effect was minimal in the present study, we believe that it was primarily due to rigorous training consisting of 24 h and the training emphasized the need to adhere to the rubric. Thus, in order to reduce measurement error attributable to raters, rigorous training as well as multiple raters should be integral part of writing assessment. Similarly, children's writing proficiency is often assessed using few tasks. Even in high stakes contexts (Olinghouse et al., 2012), for children in elementary grades (typically Grade 4), one task (e.g., Florida in 2013) or two tasks (e.g., Massachusetts in 2013) are typically used. Furthermore, many standardized writing assessments such as WIAT-3, TOWL-4, and WJ-III Writing Essay, as well as informal assessments for screening and progress monitoring progress include a single writing task. However, the present findings indicate that decisions based one or two tasks are not sufficiently reliable about children's writing proficiency, particularly when making important decisions such as state level high-stake testing or making a decision for a student's eligibility for special education services for which a high criterion reliability of .90 is applied. In these cases, even with rigorous training employed in the present study, a minimum of 4 raters and 6 tasks, or 2 raters and 7 tasks are needed for making dichotomous decisions (e.g., meet the proficiency criterion) in the narrative genre. When criterion reliability was

.80, a single rater was sufficient as long as multiple tasks were used and rater was rigorously trained. Therefore, educators in various contexts (classroom teachers, school psychologists, and personnel in state education departments) should be aware of the limitations of using a single task and a single rater in writing assessment, and use multiple tasks to the extent possible within allowable budget and time constraints.

Limitations, future directions, and conclusion

As is the case with any studies, generalizability of the current findings is limited to the population similar to the current study characteristics, including the study sample (primary grade students writing in L1), the specific measures, characteristics of raters, and the nature of training for scoring. One limitation of the present study was having different raters for different evaluative methods with an exception of productivity scoring and CBM, primarily due to practical constraints of rating a large number (approximately 1200) of writing samples. This prevented us from comparing amount of variance attributed to different evaluative methods, and it is possible that certain rater pairs may have been more reliable than others, although the rater effect was close to zero. A future study in which the same raters examine different evaluative methods should address this limitation. Another way of extending the present study is by examining the reliability of writing scores for children across grades or in different phases of writing development. As children develop writing skills, the complexity and demands of writing change, and therefore, the extent of influences of various factors (e.g., raters) might also change. Given the extremely limited number of studies with developing writers with regard to sources of variances and differences in study design in the few extant studies, we do not have concrete speculations about this hypothesis. However, it seems plausible that as ideas and sentences become more complex and dense, the influence of raters might increase in certain evaluative methods such as holistic scoring as raters' different tendencies in assigning different weights to various aspects (e.g., idea development vs. expressive language) may play a greater role in determining scores.

In addition, the order of writing tasks was not counterbalanced such that there was a potential order effect. A future replication with counterbalanced order of writing tasks is needed. Finally, it would be informative to examine the rater effect as a function of varying amount of training, particularly with classroom teachers. The present study was conducted with a specific amount of training by research team raters who were graduate students (including future school psychologists and teachers). Research assistants differ from classroom teachers in many aspects including teaching experiences and subject knowledge. Furthermore, results on holistic scoring in the present study are based on a total of 14 h of training (but 24 h across the three types of evaluations). One natural corollary is the effect of varying amount of training on reliability of different writing evaluation methods. Given that results have highly important practical implications for classroom teachers, a future study of varying intensity of training with classroom teachers would be informative.

In summary, the present study suggests that multiple factors contribute to variation in various writing scores, and therefore should be taken into consideration in writing evaluations for research and classroom instructional purposes. The present findings underscore a need to use multiple tasks to evaluate students' writing skills reliably.

Acknowledgements Funding was provided by National Institute of Child Health and Human Development (Grant No. P50HD052120). The authors wish to thank participating schools, teachers, and students.

References

Abbott, R. D., & Berninger, V. W. (1993). Structural equation modeling of relationships Among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology, 85,* 478–508.

Applebee, A. N., & Langer, J. A. (2006). *The state of writing instruction in America's schools: What existing data tell us*. Albany, NY: University at SUNY, Albany.

Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12,* 86–107.

Beck, S. W., & Jeffery, J. V. (2007). Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing, 12,* 60–79.

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

Bouwer, R., Beguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32,* 83–100.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24,* 1–21.

Brennan, R. L., Goa, X., & Colton, D. A. (1995). Generalizability analyses of work keys listening and writing tests. *Educational and Psychological Measurement, 55,* 157–176.

Coker, D. L., & Ritchey, K. D. (2010). Curriculum based measurement of writing in kindergarten and first grade: An investigation of production and qualitative scores. *Exceptional Children, 76,* 175–193.

Cooper, P. L. (1984). *The assessment of writing ability: A review of research*. GRE Board research report no. GREB 82-15R/ETS research report no. 84-12). Princeton, NJ: Educational Testing Service.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86,* 67–96.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*, 219–232.

DeVellis, R. F. (1991). *Scale development*. Newbury Park, NJ: Sage.

Duke, N. K. (2014). *Inside information: Developing powerful readers and writers of informational text through project-based instruction*. New York: Scholastic.

Duke, N. K., & Roberts, K. M. (2010). The genre-specific nature of reading comprehension. In D. Wyse, R. Andrews, & J. Hoffman (Eds.), *The Routledge international handbook of english, language and literacy teaching* (pp. 74–86). London: Routledge.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14,* 88–115.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25,* 155–185.

Espin, C. A., De La Paz, S., Scierka, B. J., & Roelofs, L. (2005). The relationship between curriculum-based measures in written expression and quality and completeness of expository writing for middle school students. *The Journal of Special Education*, *38*, 208–217.

Florida Comprehensive Assessment Test (FCAT) 2012 writing: Grade 4 narrative task anchor set. Retrieved from http://fcat.fldoe.org/pdf/G4N12WritingAnchorSet.pdf.

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review, 31*, 477–497.

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Slider, N. J., Hoffpauir, L. D., Whitmarsh, E. L., et al. (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*, 291–300.

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 for elementary school students. *School Psychology Review, 35,* 435–450.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*, 507–531.

Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology, 89,* 170–182.

Graham, S., Harris, K., & Hebert, M. (2011). *Informing writing: The benefits of formative assessment*. Washington, DC: Alliance for Excellent Education.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of the writing tasks assigned in academic degree programs*. In: TOEFL Research Report 54. Princeton, NJ: Educational Testing Service.

Hammill, D. D., & Larsen, S. C. (1996). *Test of Written Language-3*. Austin, TX: Pro-ed.

Hammill, D. D., & Larsen, S. C. (2009). *Test of Written Language-4th edition (TOWL-4)*. Austin, TX: Pro-Ed.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing, 12*, 1–9.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60,* 237–263.

Jewell J., & Malecki C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review, 34*, 27–44.

Kim, Y.-S., Al Otaiba, S., Puranik, C., Sidler, J. F., Greulich, L., & Wagner, R. K. (2011). Componential skills of beginning writing: An exploratory study. *Learning and Individual Differences, 21,* 517–525.

Kim, Y.-S., Al Otaiba, S., Sidler, J. F., & Greulich, L. (2013). Language, literacy, attentional behaviors, and instructional quality predictors of written composition for first graders. *Early Childhood Research Quarterly, 28,* 461–469.

Kim, Y.-S., Al Otaiba, S., Folsom, J. S., Greulich, L., & Puranik, C. (2014). Evaluating the dimensionality of first grade written composition. *Journal of Speech, Language, and Hearing Research, 57,* 199–211.

Kim, Y.-S., Al Otaiba, S., Wanzek, J., & Gatlin, B. (2015). Towards an understanding of dimension, predictors, and gender gaps in written composition. *Journal of Educational Psychology, 107,* 79–95.

Kondo-Brown, K. (2002). A facets analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19,* 3–31.

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31,* 329–348.

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. *Applied Measurement in Education, 2,* 195–205.

Lembke, E., Deno, S. L., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary school students. *Assessment for Effective Intervention, 28,* 23–35.

McMaster, K. L., Du, X., & Pestursdottir, A. L. (2009). Technical features of curriculum-based measures for beginning writers. *Journal of Learning Disabilities, 42,* 41–60.

McMaster, K. L., Du, X., Yeo, S., Deno, S. L., Parker, D., & Ellis, T. (2011). Curriculum-based measures of beginning writing: Technical features of the slope. *Exceptional Children, 77,* 185–206.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education, 41,* 68–84.

Moore, & T., Morton, J. (1999). *Authenticity in the IELTS academic module writing test: A comparative study of task 2 items and university assignments*. In: IELTS Research Reports No. 2 (pp. 74–116). Canberra: IELTS Australia.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavioral Research Methods, 38,* 542–547.

National Center for Education Statistics. (1999). *The NAEP 1998 writing report card for the nation and the states*, NCES 1999-462, by E. A. Greenwald, H. R. Persky, J. R. Campbell, and J. Mazzeo. Washington, DC.

National Center for Education Statistics. (2003). *The nation's report card: Writing 2002*, NCES 2003-529 by H. R. Persky, M. C. Dane, & Y. Jin. Retrieved from http://nces.ed.gov/.

National Center for Education Statistics. (2012). *The nation's report card: Writing 2011* (NCES 2012-470). Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved from http://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Authors.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw Hill.

Olinghouse, N. G. (2008). Student- and instruction-level predictors of narrative writing in third-grade students. *Reading and Writing: An Interdisciplinary Journal, 21,* 3–26.

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

Olinghouse, N. G., Santangelo, T., & Wilson, J. (2012). Examining the validity of single-occasion, single-genre, holistically scored writing assessments. In E. Van Steendam (Ed.), *Measuring writing: Recent insights into theory, methodology and practices* (pp. 55–82). Leiden: Koninklije Brill.

Puranik, C. S., Lombardino, L. J., & Altmann, L. J. (2007). Writing through retellings: An exploratory study of language-impaired and dyslexic populations. *Reading and Writing: An Interdisciplinary Journal, 20,* 251–272.

Puranik, C., Lombardino, L., & Altmann, L. (2008). Assessing the microstructure of written language using a retelling paradigm. *American Journal of Speech Language Pathology, 17,* 107–120.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22,* 1–30.

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. Van Steendam (Ed.), *Measuring writing: Recent insights into theory, methodology and practices* (pp. 1–22). Leiden: Koninklije Brill.

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

Shavelson, R., Webb, N., & Rowley, G. (1989). Generalizability theory. *American Psychologist, 44*, 922–932.

Stuhlmann, J., Daniel, C., Delinger, A., Denny, R. K., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Journal of Reading Psychology, 20,* 107–127.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., de Kruif, R. E. L., Reed, M., et al. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Education and Psychological Measurement, 59,* 492–506.

Tillema, M., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2012). Quantifying the quality difference between L1 and L2 essays: A rating procedure with bilingual raters and L1 and L2 benchmark essays. *Language Testing, 30,* 1–27.

van den Bergh, H., De Maeyer, S., van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. In E. Van Steendam (Ed.), *Measuring writing: Recent insights into theory, methodology and practices* (pp. 23–32). Leiden: Koninklije Brill.

Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. *Reading and Writing: An Interdisciplinary Journal, 24,* 203–220.

Wechsler, D. (2009). *Wechsler Individual Achievement Test-3rd edition (WIAT-3)*. San Antonio, TX: Pearson.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15,* 263–287.