Reading and Writing
https://doi.org/10.1007/s11145-020-10053-1

Exploring the dimensionality of kindergarten written composition

Cynthia Puranik[1] · **Molly Duncan**[1] · **Hongli Li**[1] · **Guo Ying**[2]

Abstract
Despite increasing pressure for children to learn to write at younger ages, there are many unanswered questions about composition skills in early elementary school. The goal of this research was to examine the dimensionality of composition skills in kindergarten children, thereby adding to current knowledge about the measurement of young children's writing and its component skills. The writing of 282 kindergarten children were assessed using three different scoring methods. Confirmatory factor analyses were used to investigate the dimensionality of various methods of scoring. Results indicated that a qualitative scoring system and a productivity scoring system capture distinct dimensions of kindergartners' compositions. A scoring system for curriculum-based measurement could not attain acceptable fit, which may suggest that CBM is ill-suited for capturing the important components of composition for kindergartners. This study indicated that the measurement and components of composition in kindergarten may be qualitatively different from the compositions of older children.

Keywords Component skills · Confirmatory factor analysis · Dimensionality of writing · Early writing · Kindergarten · Writing assessment

Introduction

The measurement of composition skills has grown increasingly important for educators and researchers due to pressure for children to write at young ages and an increased emphasis on data-driven decision making in the classroom. Despite the important role of writing in academic learning, there are still many open questions and a lack of consensus about how best to measure young children's

✉ Cynthia Puranik
cpuranik@gsu.edu

[1] Department of Communication Sciences and Disorders, Georgia State University, Ste 850, 30 Pryor St. SW, Atlanta, GA 30303, USA

[2] University of Cincinnati, Cincinnati, USA

Published online: 05 May 2020

composition. Composition refers to children's ability to generate ideas for what to write and compose phrases, sentences, or texts in their writing using conventional or invented spelling. One important issue revolves around how to best capture children's written output, or in other words, what components of written composition are important. Some studies have reported or assumed that young children's composition is accurately represented as a single, holistic component or score (e.g., Abbott & Berninger, 1993; Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006), whereas others have found that the writing of elementary school students contains many components, including macro-organization, productivity, complexity, and accuracy (Hall-Mills & Apel, 2015; Kim, Al Otaiba, Folsom, Greulich, & Puranik, 2014, Kim, Al Otaiba, Wanzek, & Gatlin, 2015; Puranik, Lombardino, & Altmann, 2008; Wagner et al., 2011).

A second issue is what type of measurement best captures these important components of writing. Researchers have used many different methods to measure elementary school children's compositions, including qualitative scoring systems, quantitative scoring, and curriculum-based measures (CBM; Abbott & Berninger, 1993; Dockrell, Ricketts, Charman, & Lindsay, 2014; McMaster & Espin, 2007; Puranik & Al Otaiba, 2012; Wagner et al., 2011). These various scoring methods may be capturing different components of composition ability. Consequently, there is little consensus regarding what composition is and how it is best assessed during early childhood. A greater understanding of the measurement of young children's composition is necessary so that researchers can continue to appropriately assess writing, examine predictors of composition ability, and design interventions for struggling composers. A greater understanding of measurement will also enable more accurate identification of students who are struggling and may allow practitioners to pinpoint the specific skills that must be supported.

The primary aim of this study was to examine the dimensionality of writing in kindergarten children. To date, there have been no studies examining the dimensionality of composition in kindergarten children. Kindergarten is a year when children are just learning to write, including learning to write letters, spell words, and write sentences. Because they are just beginning to learn to write, their compositions may be qualitatively and quantitatively different than the compositions of older children. This study attempted to add to the existing body of literature by replicating previous findings about dimensionality in older children, examining the dimensionality of a less-studied qualitative scoring system, examining alternate possibilities of the dimensionality of CBM, and extending the research to kindergarten compositions.

Components of young children's compositions

Previously identified dimensions of composition include macro-organization, accuracy, productivity, and complexity. The findings of recent studies regarding the dimensions present in the writing of young children and the indicators of each dimension are discussed below.

Macro-organization

The macro-organization of writing is generally considered to be the most important outcome or measurement of written composition (e.g., Kim, Al Otaiba, Wanzek, et al., 2015), perhaps because the writing of older students is judged primarily on its organization and content (e.g., ACT, 2018). The macro-organization component of writing is typically defined as the content of the ideas and the overall organization of the composition (Hall-Mills & Apel, 2015; Kim et al., 2014; Wagner et al., 2011), though some researchers include components such as word choice and sentence fluency (e.g. Kim et al., 2014).

Accuracy

The accuracy of the use of writing conventions is sometimes considered as a distinct component of writing. This component typically includes the accuracy of spelling, grammar, and mechanics such as punctuation and capitalization (Kim et al., 2014). This is also a component of composition that is considered in many standardized tests. For example, the Georgia Standards of Excellence where part of this research was conducted require kindergartners to "demonstrate command of the conventions of standard English grammar and usage when writing or speaking" and "demonstrate command of the conventions of standard English capitalization, punctuation, and spelling when writing," (Georgia Department of Education, 2015, p. 5).

Productivity

Because of the difficulty of reliably rating the quality of young children's writing, a common approach is to use writing productivity as an outcome for kindergartners and first graders (Kent, Wanzek, Petscher, Al Otaiba, & Kim, 2014; Kim et al., 2011; Puranik & Al Otaiba, 2012; Puranik, Al Otaiba, Sidler, & Greulich, 2014). Writing productivity is measured by counting the number of words, different words, ideas, clauses, or sentences (Berninger et al., 1992; Graham, Berninger, Abbott, Abbott, & Whitaker, 1997; Kent et al., 2014; Kim, Park, & Park, 2013; Puranik et al., 2008). For young composers, writing quality and productivity are related both conceptually (Kim, Al Otaiba, Wanzek, et al., 2015) and empirically (Abbott & Berninger, 1993; Kim et al., 2014; Nelson & Van Meter, 2007), but are nevertheless distinct constructs (Kim et al., 2014; Kim, Al Otaiba, Wanzek, et al., 2015; Wagner et al., 2011). Their conceptual link stems from the fact that children who write greater quantities of text have more opportunities to convey complex, meaningful ideas.

Syntactic complexity

Syntactic complexity is another distinct component of children's writing (Kim et al., 2014; Puranik et al., 2008; Wagner et al., 2011). There are multiple methods for scoring syntactic complexity, but the most common methods rely on assessing the number or ratio of main and subordinate clauses in the composition. This may be a valuable component of writing because writers are expected to produce

compositions that include a variety of sentence structures. Previous studies have identified syntactic complexity as a separate dimension of children's writing, albeit with slightly older children from 1st grade and above. The majority of the essays in this study contain either no complete clauses or only one. Since little variation is expected in these scores, the present study did not measure syntactic complexity.

Methods for measuring young children's compositions

Ideas on how best to measure the important components of children's writing vary. Historically, researchers often used a single score to capture the quality of written composition, though researchers vary in their definition of quality (e.g. Abbott & Berninger, 1993; Berninger et al., 1992; Graham et al., 1997; Olinghouse, 2008). More recently, researchers have turned to other analytical scoring methods such as quantitative scoring and CBM. These scoring systems, along with the components of composition they are hypothesized to capture, are discussed below.

Qualitative scoring systems

Several scoring systems have been developed to measure the quality of writing. One such scoring system is the 6 + 1 Traits rubric (Education Northwest, 2017). The 6 + 1 Traits Rubrics system is widely used, freely available, and frequently researched (e.g. Gansle et al., 2006; Kim et al., 2014), however, there is limited data on its technical adequacy and scoring reliability. The rubric contains seven different categories that each have criteria for scoring multiple aspects within that category. For example, the organization category contains criteria for scoring the quality of the composition's beginning, middle, and end; transitions; sequencing; and title. Although there are seven different categories, research indicates that the rubric captures two distinct dimensions of writing for first graders: scores for ideas, organization, word choice, and sentence fluency capture the macro-organization of the writing, whereas the spelling, mechanics, and handwriting categories capture the technical accuracy of the writing (Kim et al., 2014).

Coker and Ritchey (2010) have proposed a similar but pared-down scoring system for scoring the quality of short writing samples from children as young as kindergarten. The categories of the scoring system were selected to reflect important features of writing that are reasonable expectations for young writers. Although it has not been studied as extensively as other methods of scoring, it shows promise as a quick and reliable measure of writing with acceptable criterion-related validity compared to more established measures (Coker & Ritchey). Similar to the 6 + 1 Traits Rubric, this scoring has five categories.

Response type The response type category measures the completeness and complexity of composition. In kindergarten, children are typically graduating from writing single letters at a time to writing entire sentences, so the length and complexity of a short composition is a developmentally appropriate and sensitive measure of writing ability (Berninger, Fuller, & Whitaker, 1996; Coker & Ritchey,

2010). The inclusion of response type as a category is in line with the 6 + 1 Traits Rubrics because the sentence fluency category awards points for using a greater variety of sentences and more complex sentence structures.

Relationship to prompt The relationship to prompt category measures whether the composition is related to the topic and how much the child elaborates on that topic. This is a key element for the macro-organization of the text, and variations of this category are present in nearly every qualitative writing scoring system. For example, in the 6 + 1 Traits Rubrics, children receive points in the ideas category for including a clear main idea with supporting details. Under the scoring system used by Hall-Mills and her colleagues, the organization category awarded points when compositions included a clear beginning and supporting details (Hall-Mills, 2010; Hall-Mills & Apel, 2015). The scoring system used by Wagner et al. (2011) conceptualized this category slightly differently by awarding points based on the inclusion of a topic sentence and the number of key elements (main idea, body, and conclusion). Wagner and colleagues' scoring system was appropriate for their age group (1st and 4th graders) and their prompt.

Grammatical structure The grammatical structure category measures how many grammatical mistakes a child makes and how those mistakes impact the meaning of the sentence. Grammatical accuracy is a consideration in most measurement systems. Most researchers group this measure with the accuracy of writing (Education Northwest, 2017; Puranik et al., 2008) or include it as a unique dimension (Hall-Mills & Apel, 2015).

Spelling Measurements of spelling accuracy are also included in nearly every writing scoring system. Kim et al. (2014) found that this measure fit on a factor that represented the accuracy of writing conventions.

Mechanics Mechanical accuracy is also frequently measured in writing. Rating systems designed for more mature writers sometimes include more stringent criteria, such as correct capitalization of proper nouns and titles (e.g., Education Northwest, 2017) whereas rating systems for younger less mature writers might include only capitalization of first letter in the sentence and correct ending punctuation (Puranik et al., 2008; Wagner et al., 2011).

Productivity scoring system

There is more agreement regarding the variables that should be included in productivity scoring systems. The most frequently included variable is words written (WW), others include the number of ideas, the number of different words, and the number of minimal terminable units.

Words written (WW) Although different systems of productivity measurement vary in which units they count, a measure of WW is included in nearly all of them (Hall-Mills & Apel, 2015; Kim et al., 2014; Kim, Al Otaiba, Wanzek, et al., 2015; Puranik et al., 2008; Tindal & Parker, 1989; Wagner et al., 2011). Under this system, words that are repeated throughout the composition are counted each time they appear.

Number of ideas The number of ideas in a composition is the number of complete propositions, or subject-predicate pairs (e.g. Kim et al., 2011, 2014; Puranik et al., 2008). This is an important supplement to the WW measure because it awards credit to writers who express complex ideas concisely. These writers might be recognized as less productive if WW is the only measure of productivity.

Curriculum-based measures

Demands for accountability and quantifiable learning in education have led educators and researchers to develop measures of academic skills that are easy and quick to administer, that can be scored by teachers, and that can track students' growth; these measures are commonly called curriculum-based measures (CBM; Hosp, Hosp, & Howell, 2007). CBM for writing involves scoring a short composition according to the correct or incorrect word sequences (CWS, IWS).[1] A word sequence refers to a pair of two consecutive words or a consecutive word and punctuation mark. A correct word sequence (CWS) is a pair that is both contextually and grammatically correct (Hosp et al., 2007). Some have argued that CBM scores capture both production-dependent and production-independent aspects (Tindal & Parker, 1989).

Correct word sequences (CWS) Although the number of CWS is related to the accuracy of the writing conventions (for example, the number of words a child spells correctly partly determines the CWS score), Tindal and Parker (1989) have demonstrated that it is a production-dependent measure. In this sense, CWS may be considered a measure of a writing productivity. Conversely, it may capture the writing fluency component hypothesized by Kim and colleagues (Kim, Al Otaiba, Wanzek, et al., 2015; Kim, Gatlin, Al Otaiba, & Wanzek, 2018).

Percent of correct word sequences (%CWS) Many researchers choose to use scores derived from CWS rather than the raw scores themselves. For example, researchers have used the number of correct minus incorrect word sequences or the percent of correct word sequences (%CWS) out of total word sequences (McMaster & Espin, 2007). Tindal and Parker (1989) considered %CWS to be a production-independent measure, conceptually distinct from CWS and other productivity measures. They

[1] When CBM is administered as the only measure of composition, it sometimes includes a count of the WW and a count of the words spelled correctly or incorrectly. However, these measures are already accounted for in the quality and productivity scoring systems used in this study.

also found that %CWS was an indicator of the writing quality score, which they defined as a holistic score that captured "communicative effectiveness" (p. 175), and that comprised both macro-organization and technical accuracy. Because %CWS is determined primarily by technical accuracy (such as the correct spelling of two consecutive words, or the correct usage of punctuation), it may be an indicator of technical accuracy. Conversely, Kim and colleagues included %CWS in their dissociable CBM or writing fluency measure (Kim, Al Otaiba, Wanzek, et al., 2015; Kim et al., 2018), separate from the macro-organization and productivity measures. Therefore, it is unclear whether %CWS is an additional measure of the technical accuracy of writing, or whether %CWS (together with CWS) captures a distinct component of writing that is dissociable from components such as accuracy and productivity. This study tested both possibilities.

The present study

The present study used essays written by kindergartners near the end of the school year to investigate how many dimensions exist when using multiple evaluation approaches/methods of scoring. The first research question examined how many dimensions are present in kindergarten compositions when a qualitative scoring system is used. For this study, we used a modified version of Coker and Ritchey's (2010) quality sentence scoring rubric to rate compositions, as it has been shown to be developmentally appropriate for kindergarten children. As mentioned previously, this qualitative scoring rubric awards points in five categories: response type, relationship to prompt, grammatical structure, spelling and mechanics. However, research is needed to determine its dimensionality. Like the more complex 6 + 1 Traits Rubrics, its five categories may be separable into two distinct dimensions. Response type, relationship to prompt, and grammatical structure may capture the macro-organization of the composition, whereas spelling and mechanics may both measure aspects of the accuracy of writing conventions. The present study examined whether this qualitative scoring system captures two distinct dimensions of writing (macro-organization and accuracy) or a single dimension (which may represent the overall quality of the writing). It was hypothesized that the qualitative scores would capture two distinct dimensions of writing in line with the findings of Kim et al. (2014) with first graders. An alternate possibility was that the scoring system would be best represented as a single factor that may capture the overall quality of the writing.

In ratings for older students, the macro-organization factor generally includes inclusion of a topic sentence or key elements (e.g. story elements like plot, character, and setting). However, this may be too stringent of an expectation for kindergartners, who often write only a sentence or less. Therefore, in the hypothesized two-dimensional model for kindergarteners, response type and relationship to prompt was expected to be related to macro-organization because these two categories are measuring a similar concept to what is measured in scoring systems for more mature writers. The third category expected to be related to

macro-organization was grammatical structure. This is in contrast to the more conventional approach of including it with the measures of accuracy. Coker and Ritchey (2010) argue that severe mistakes or having many mistakes can compromise the meaning of the composition or make it impossible to decipher (see also Olinghouse, 2008). This is particularly true for young writers, who tend to have a high percentage of grammatical mistakes, and who tend to include much less context that can clear up confusion. This inherent link with the meaning of the composition, which in some ways is unique to the age group in this study, may make it a good fit for the macro-organization factor. Finally, spelling and mechanics were expected to measure the accuracy of writing conventions, as it does in most other rating systems (e.g. Kim et al., 2014; Wagner et al., 2011).

The second research question examined how many dimensions are present in kindergarten compositions when a productivity scoring system is used in addition to the qualitative scoring system. It was hypothesized that the productivity scoring system would capture a dimension of writing that was separate from the dimension captured by the qualitative scoring system. This would be consistent with previous studies of children's writing (e.g., Kim et al., 2014, Puranik et al., 2008, Wagner et al., 2011). An alternate possibility was that the productivity scoring system would capture the same dimension as the qualitative scoring system. In line with a great deal of previous research (e.g. Kim et al., 2014; Puranik et al., 2008; Wagner et al., 2011), the WW and Ideas measures were hypothesized to load onto a productivity factor that would be distinct from the factors of the qualitative scoring system.

The third research question examined how many dimensions are present in kindergarten compositions when CBM is used for scoring. Although CBM has many useful properties, including the fact that it is quick and reliable to score and can capture growth across a school year, it is still unclear exactly which aspects of writing CBM captures. Examining the dimensionality of two of its measures may help to illuminate exactly what educators and researchers measure when they use CBM. This is an important consideration, given its prevalence in research and in recommendations for educators (e.g. Deno, 2003; Hosp et al., 2007; Kim, Al Otaiba, Wanzek, et al., 2015; McMaster et al., 2011; McMaster & Espin, 2007; Tindal & Parker, 1989).

One possibility is that the two CBM scores would capture separate dimensions of writing. These dimensions may represent productivity and accuracy, in line with Tindal and Parker's (1989) findings of production-dependent and production-independent dimensions. Another possibility was that together, the CBM scores would capture a single dimension of writing (Kim, Al Otaiba, Wanzek, et al., 2015; Kim et al., 2018). Because of conflicting findings from previous studies, there was no a priori hypothesis about which of the two models would fit better.

The fourth research question examined how many dimensions are present in kindergarten compositions when a qualitative scoring system, a productivity scoring system, and CBM are used for scoring. One possibility was that %CWS would capture the technical accuracy of the writing, whereas CWS would capture the productivity of the writing, resulting in a two-factor model. An alternate possibility was that the two CBM scores together would capture a dimension of writing that

is distinct from the other dimensions in this study, resulting in a three-factor model. There was no a priori hypothesis about which of the two models would fit better.

Methods

Participants

The participants in this study were 281 kindergarten students recruited from public schools serving urban and suburban neighborhoods in the South and Midwest United States.

These students attended 49 different classrooms, with each classroom having on average six participating students. There was one additional child who participated in data collection but refused one of the essays. That child's scores were not used for any of the analyses in this paper.

Ninety seven percent (273) of the students' parents returned a questionnaire containing demographic information. The average child age was about 6.1 years (range 5.6–7.0 years). Other demographic information is presented in Table 1.

Measures and procedures

Human subjects' approval was obtained from the Institutional Review Board prior to conducting this research. Kindergarten children wrote two essays on two separate days near the end of the school year (April or May). Writing took place in a convenient location at the child's school, usually in a group of about six children. In the first prompt, the examiner instructed children to write about a special event (essay 1). The examiner introduced the writing topic using a script, saying "Today, you are going to draw and write about a special event in your life." The script the examiner followed gave examples of special events (a special birthday or a special vacation), elicited an idea from each child, and asked for an additional detail from each child (such as "Who was there?"). Then the examiner also instructed children to try to keep writing for the entire time, to sound out words as best they could, and to cross out mistakes instead of erasing them. A different prompt was used for the second essay, although the structure of the instructions remained the same. In the second prompt, the examiner instructed children to write about something they were an expert on or knew a lot about (essay 2). The examples given were about topics that children might know a lot about such as lions, cars, or dinosaurs. An example of an additional detail the examiner asked for was, "What does it/they look like?".

Children spent 5 min independently drawing a picture and writing their essay. During the 5 min, they were not given any assistance with writing (including spelling). After 5 min, examiners asked children to read what they had written so that the examiner could write it in the margins of the paper. This aided scoring in cases where children had poor handwriting or spelling.

Table 1 Demographic characteristics of sample

Variable	Responses	N	Percent
Gender	Male	128	46.9
	Female	144	52.7
	No response	1	.4
Ethnicity	Black/African American	57	20.9
	White/Caucasian	188	68.9
	Hispanic or Latino	4	1.5
	Asian (Indian)	4	1.5
	Biracial or Multiracial	12	4.4
	Other	2	.7
	No Response	3	1.1
Child home language	English only	267	97.8
	English and other language	4	1.5
	Other language Only	1	.4
	No response	1	.4
Highest education of mother	Less than high school diploma	14	5
	High school diploma	40	14.7
	Post high-school training	48	17.6
	Two-year degree	12	4.4
	Four-year degree	84	30.8
	Graduate degree	74	27.1
	No response	1	.4
Annual family income	$20,000 or less	33	12.1
	$20,001–$40,000	41	15.0
	$40,001–$60,000	20	7.3
	$60,001–$85,000	38	13.9
	$85,001 or more	137	50.2
	No response	4	1.5

Essay scoring

Each essay was scored for quality (qualitative indicators, including response type, relationship to prompt, grammatical structure, spelling, and mechanics), productivity (WW and ideas) and CBM (CWS and %CWS). Coker and Ritchey's (2010) original scoring system was slightly modified to fit the different task requirements and prompts used in this study. The first modification was related to the grammatical structure category and was necessary due to the length of the writing samples in this study. Coker and Ritchey asked participants to write two sentences about a prompt, whereas in this study children wrote longer compositions. Coker and Ritchey's grammatical structure category awards two points to sentences that contain a single grammatical error and one point to sentences with more than one error. Since the participants in this study sometimes wrote longer compositions, the scoring

system in this study allowed two points for compositions even if they contained multiple grammatical errors, provided that the errors did not comprise more than 50% of the writing sample and did not have a major effect on the meaning. The second major modification was of the relationship to prompt category. Coker and Ritchey's scoring system was designed to score sentences about any prompt. However, in this study, the scoring systems needed to apply to only two topics, and needed to capture differences in slightly longer compositions. Therefore, the scoring system used in this study maintains the spirit of Coker and Ritchey's scoring system by awarding points for details that are appropriately related to the prompt, while being more specific to the respective prompts to maximize scoring reliability.

Inter-rater reliability Four graduate research assistants (GRAs) were trained extensively by the first author on kindergarten essays from a previous study until they reached a reliability of 80%. For scoring the essays in this study, GRAs worked in pairs to score the essays for quality (response type, relationship to prompt, grammatical structure, spelling, and mechanics), productivity (WW and ideas), and CBM (CWS and %CWS). Each assessment was individually scored by two GRAs, and reliability was calculated for each measure. Because the qualitative scoring system had four possible scores for each category, it was treated as an ordinal measure and reliability was measured with Cohen's kappa. Interrater reliability ranged from .71 to .96. Because productivity (WW and ideas) and CBM (CWS and %CWS) scores are continuous measures, reliability was measured by intraclass correlation coefficient (ICC); reliability ranged from .92 to 1.0. Following the reliability calculations, each GRA pair compared scores and came to an agreement about any discrepancies before recording the final score.

Analytic strategy

Preliminary statistics for this analysis (such as normality tests and correlations) were conducted in RStudio (RStudio Team, 2016). Modeling analyses were performed with Mplus, version 8.1 (Muthén & Muthén, 2017). Given that students were nested within classrooms, all our analyses accounted for the nested nature of the data using cluster-corrected standard errors.

Most of the analyses presented in this paper contain some ordinal indicators, specifically the scores from individual categories of the qualitative scoring system. Traditional maximum likelihood (ML) estimation does not perform well in confirmatory factor analysis (CFA) with ordinal data, so weighted least squares means and variances (WLSMV) estimation was used for most of the analyses, as recommended by Finney and DiStefano (2013) and Bandalos (2014). When used with ordinal data with four categories, WLSMV is more likely to result in unbiased parameter estimates compared to ML or robust ML estimation (Bandalos). The only exception was the third research question, which contained only continuous indicators and therefore did not require WLSMV. Instead, the analyses for Question 3 used robust maximum likelihood estimation (MLR). MLR is also robust to nonnormal data (Brown, 2015), and this was important for this study because skewness and kurtosis tests

(performed in the moments package of R; Komsta & Novomestky, 2015) were significant for several variables.

For each question, model fit was examined with model-fit statistics and Chi square tests of difference (for questions with categorical data, Chi square tests were performed with the DIFFTEST option in Mplus; Muthén & Muthén, 2017). The Chi square test was important for determining the relative fit of models, that is, which model fit better than another. However, Chi square values are sensitive to sample size (Kline, 2016; Marsh, Hau, Balla, & Grayson, 1998). In addition, CFI, TLI, and RMSEA were used to evaluate model fit. Because the measurement of complex skills such as writing is somewhat unreliable by nature for kindergartners (e.g., McMaster & Espin, 2007), less conservative rules of thumb were employed for determining reasonable model fit. Specifically, values of about .90 or higher for CFI and TLI and about .10 or lower for RMSEA were deemed reasonable (Browne & Cudeck, 1993; Hu & Bentler, 1999). All the initial models had poor fit, so after determining whether the more parsimonious or less parsimonious model fit better based on theory, modification indices were also examined to highlight areas with poor local fit and suggest improvements (Brown, 2015). When the modification indices suggested freely estimating parameters that were theoretically sensible, these parameters were added one at a time to the better-fitting model.

There were two main types of parameter additions that made sense theoretically. The first was correlation between the errors of two indicators from the same essay (for example, relationship to prompt and response type for Essay 1). It is reasonable to think that because these indicators were based on a single essay, their error variances would be related. The second theoretically sensible correlation was between the errors of the same measure for different essays (for example, mechanics for Essay 1 with mechanics for Essay 2). This is another reasonable suggestion, because children who use good or poor punctuation or capitalization in one essay are likely to do so on the second.

Results

Descriptive statistics

Means and standard deviations for each measure are presented in Table 2. Descriptive data indicated that on average, children wrote nine words for the special event essay and about eight words for the expert essay. The CBM scores indicated a great deal of variability in both CWS and %CWS. Due to the fact that neither of these variables could be lower than 0, both had strong positive skews; however, the estimation methods used in the CFAs of this study are capable of handling non-normal data. Examination of the pattern of scores on the qualitative scoring systems did not reveal any distinct patterns whereby one category was easier in one essay than the other.

Correlations between the observed variables are presented in Table 3. Correlations between two continuous variables were Pearson correlations; correlations between a continuous and a categorical variable or between two categorical

Table 2 Descriptive statistics for composition measures

Measure	Essay 1 (special event)		Essay 2 (expert)	
	Score	Score frequency	Score	Score frequency
Response type	0	25	0	35
	1	45	1	38
	2	135	2	145
	3	76	3	63
Relationship to prompt	0	129	0	133
	1	63	1	78
	2	42	2	38
	3	47	3	32
Grammatical structure	0	61	0	63
	1	19	1	31
	2	106	2	104
	3	95	3	83
Spelling	0	31	0	39
	1	76	1	89
	2	158	2	137
	3	16	3	16
Mechanics	0	50	0	72
	1	140	1	129
	2	70	2	67
	3	21	3	13
Measure	Mean	*SD*	Mean	*SD*
Words written	9.00	6.13	8.11	5.83
Ideas	1.36	1.23	1.32	1.17
CWS	3.90	4.36	3.04	4.13
%CWS	33.65	25.18	27.48	24.42

$N = 281$ for all scores for both essays

variables were Spearman correlations, as this type of correlation is more appropriate for categorical data (Rugg, 2007). With a few exceptions that are discussed in more detail below, correlations were small to moderate.

Dimensions in the qualitative scoring system

Two alternative CFA models were fit to test the dimensionality of the qualitative scoring system: a one-factor model, in which the scoring system was unidimensional with all five indicators for both essays (10 total indicators) loading onto a single factor, and a two-factor model with two dimensions: macro-organization (relationship to prompt, response type, and grammatical structure) and accuracy of writing conventions (spelling and mechanics). The fit for both models was poor, therefore

Table 3 Correlations between observed variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. Response Type 1	–									
2. Response Type 2	.45	–								
3. Rel. to Prompt 1	.51	.33	–							
4. Rel. to Prompt 2	.33	.53	.37	–						
5. Gram. Structure 1	.57	.38	.34	.31	–					
6. Gram. Structure 2	.27	.65	.25	.41	.34	–				
7. Spelling 1	.42	.33	.27	.28	.43	.29	–			
8. Spelling 2	.31	.45	.16	.25	.31	.41	.39	–		
9. Mechanics 1	.51	.40	.28	.23	.42	.20	.42	.28	–	
10. Mechanics 2	.34	.56	.19	.34	.26	.39	.32	.42	.50	–
11. Words Written 1	.75	.41	.51	.32	.32	.18	.38	.28	.38	.30
12. Words Written 2	.45	.69	.35	.46	.24	.30	.30	.40	.30	.43
13. Ideas 1	.84	.43	.46	.31	.48	.24	.37	.27	.45	.35
14. Ideas 2	.48	.82	.37	.55	.28	.55	.30	.38	.36	.45
15. CWS 1	.63	.43	.36	.34	.44	.27	.70	.43	.58	.43
16. CWS 2	.42	.60	.24	.35	.36	.47	.46	.71	.39	.61
17. %CWS 1	.39	.33	.16	.28	.44	.27	.75	.41	.54	.38
18. %CWS 2	.34	.47	.15	.25	.34	.51	.43	.74	.33	.60
	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)		
11. Words Written 1	**–**									
12. Words Written 2	**.58**	**–**								
13. Ideas 1	**.83**	**.54**	**–**							
14. Ideas 2	**.50**	**.80**	**.54**	**–**						
15. CWS 1	**.74**	**.52**	**.66**	**.50**	**–**					
16. CWS 2	**.43**	**.65**	**.42**	**.62**	**.61**	**–**				
17. %CWS 1	**.29**	**.28**	**.29**	**.28**	**.73**	**.48**	**–**			
18. %CWS 2	**.27**	**.34**	**.25**	**.37**	**.49**	**.79**	**.56**	**–**		

Correlations in boldface are Pearson correlations; standard print are Spearman correlations. All correlations are significant, $p < .01$

Rel. to Prompt relationship to prompt, *Gram. Structure* grammatical structure, *CWS* correct word sequences, *CIWS* correct minus incorrect word sequences; %CWS percent correct word sequences

we decided to modify the model based on modification indices and theoretical considerations. There were three important theoretical considerations. First, previous research has shown that young children's text generation is significantly constrained by their transcription skills (Graham et al., 1997; Puranik & Al Otaiba, 2012). Second, Coker and Ritchey (2010) reported that the sentence writing quality score taps a unitary dimension in kindergarten students' written performance. Finally, we calculated Cronbach's alpha for the quality score of the two essays to investigate the degree to which items assessed a single construct. Internal consistency reliability for Essay 1 was .99 and for Essay 2 was .82.

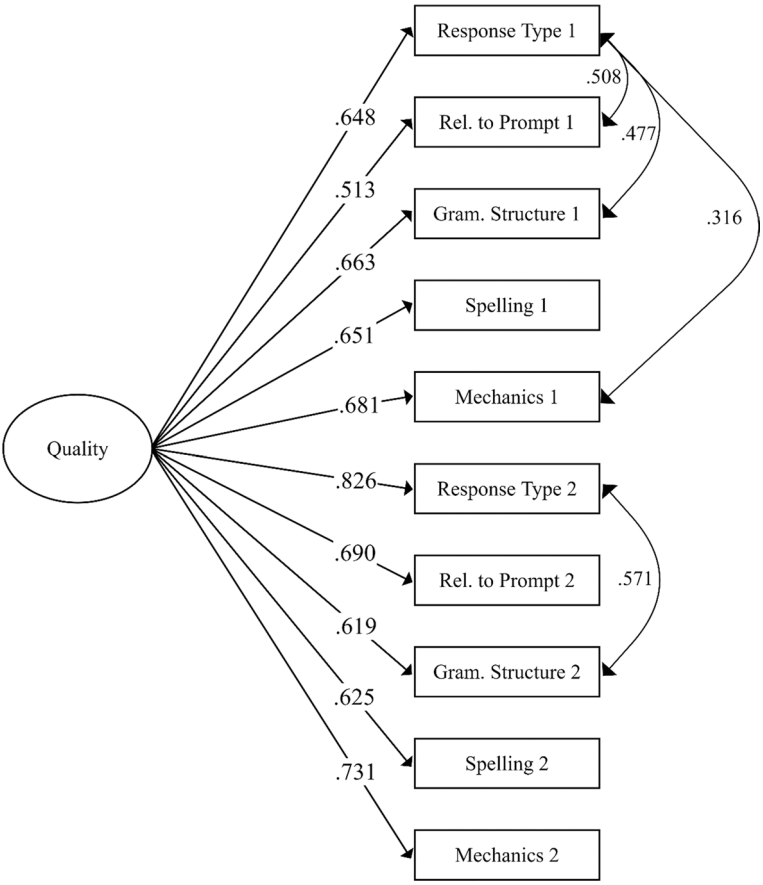

Fig. 1 Standardized factor loadings of the final model for the qualitative scoring system. *Gram. Structure* grammatical structure, *Rel. to Prompt* relationship to prompt

Adding the modifications suggested in the modification indices rapidly increased the correlation between the two factors, and many of the modification indices suggested cross-loading indicators on both factors. For these reasons, it seemed preferable to retain the one-factor model and use the modification indices and theoretical considerations to improve it. The final one-factor model for the qualitative scoring system is depicted in Fig. 1, and model fit statistics are presented in Table 4: $\chi^2(31) = 137.071$, CFI = .951, TLI = .929, RMSEA = .11 (.092, .129).

Dimensionality of quality and productivity

The second research question examined whether the qualitative scoring system and the productivity scoring system represent a single dimension or distinct dimensions of kindergarten composition. Preliminary data screening revealed a problematically high correlation between the response type indicator and the

Table 4 Overall model fit indices for final models

Research question	Figure number	Model description	χ^2 (*df*, *p*)	CFI (TLI)	RMSEA (90% confidence intervals)
1	–	Initial one-factor model	226.757 (35, < .001)	.911 (.886)	.139 (.122–.157)
1	–	Initial two-factor model	205.408 (34, < .001)	.920 (.895)	.134 (.116–.152)
1	1	Final one-factor model	137.071 (31, < .001)	.951 (.929)	.110 (.092–.129)
2	–	Initial one-factor model	308.862 (54, < .001)	.798 (.753)	.129 (.116–.144)
2	–	Initial two-factor model	236.661 (53, < .001)	.855 (.819)	.111 (.097–.125)
2	2	Final two-factor model	170.554 (50, < .001)	.905 (.874)	.092 (.077–.108)
3	–	One-factor model (final)	40.090 (2, < .001)	.829 (.487)	.260 (.193–.333)
3	–	Two-factor model	37.740 (1, < .001)	.835 (.010)	.361 (.268–.464)
4	–	Initial two-factor model	596.936 (103, < .001)	.732 (.687)	.130 (.120–.141)
4	–	Initial three-factor model	596.795 (101, < .001)	.731 (.680)	.132 (.122–.142)
4	3	Final two-factor model	331.231 (90, < .001)	.869 (.825)	.097 (.086–.109)

CFI comparative fit index, *RMSEA* root mean square error of approximation

ideas indicator for each essay, above $r = .80$ in both cases. In theory, these two measures are closely related but not identical. Response type captures the completeness of the response, with one point awarded for having one to several words and up to three points awarded for multiple sentences or a complex sentence. Ideas is a measure of how many complete propositions exist in the writing. Therefore, response type is a more lenient indicator in that it awards points for a lower standard (such as a few words that don't make a complete sentence); however, it has a maximum of three points. Thus, a composition with several complete sentences would receive the same score as a composition with two complete sentences. Conversely, the ideas measure does not award points for incomplete sentences, but it can award a theoretically infinite amount of points for compositions with more complete propositions. With our sample, these two measures were practically identical. There were many compositions that contained a few words but no complete sentences. However, there were few that exceeded two complete sentences or a single complex sentence. Thus, there was not sufficient variation at the higher end of the spectrum to make ideas a distinct indicator.

The close relationship between these two variables resulted in difficulty in model convergence. When variables are too highly correlated, it is best to

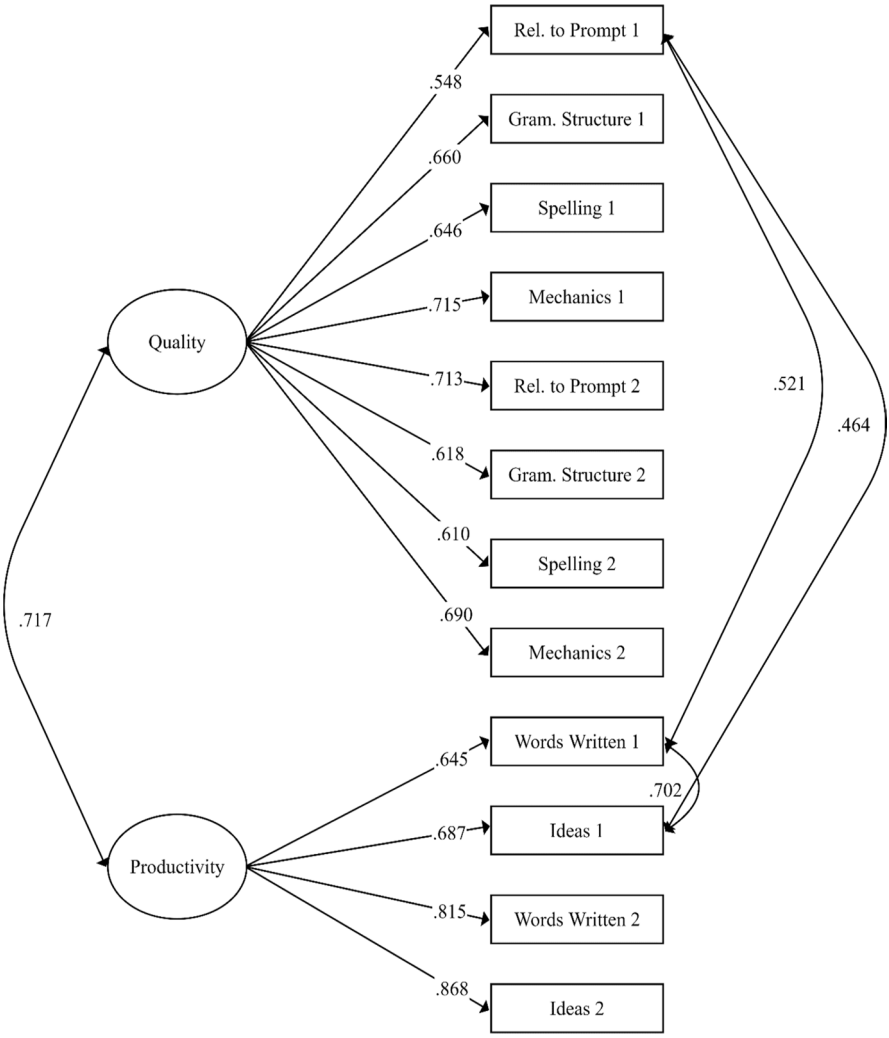

Fig. 2 Standardized factor loadings of the final model for the qualitative scoring system with productivity indicators. *Gram. Structure* grammatical structure

combine or drop one of them (Kline, 2016). Response type was dropped from the models because the ideas indicator is more widely represented in writing research (e.g. Kim et al., 2014; Puranik et al., 2008; Wagner et al., 2011) than the response type indicator. Furthermore, it seemed better to retain the indicator that had a larger possible range of values. Additionally, dropping the ideas indicators would have resulted in only two indicators (the WW indicators) loading onto the productivity factor. Two-indicator factors can be problematic for identification, and they can be problematic because they allow more measurement error (Kline, 2016).

The model in which productivity was a unique factor had significantly better fit than the one-factor model, $\chi^2(1) = 91.28$, $p < .001$, but the fit of both models was poor. The addition of several theoretically sensible correlated error terms that were suggested by the modification indices improved the model fit to acceptability. The final model with standardized factor loadings is depicted in Fig. 2, and model fit statistics are presented in Table 4: $\chi^2(50) = 170.554$, CFI = .905, TLI = .874, RMSEA = .092 (.077, .108).

Dimensionality of CBM indicators

The purpose of Research Question 3 was to examine the dimensionality of the CBM indicators when they are used independently of other measures. This analysis compared the fit of a unidimensional and a two-dimensional model. The fit for both models was unacceptably poor. Modifications to improve the model fit were not attempted because the model had only two degrees of freedom. Model fit statistics are presented in Table 4.

Dimensionality of all composition measures

The purpose of Research Question 4 was to build on the findings of the previous models and additionally determine the best model for accommodating the CBM indicators. Two models were fit. In the first model, CWS (a production-dependent CBM measure) loaded onto the productivity factor, along with WW and ideas, whereas %CWS loaded onto the writing quality factor, along with qualitative scoring system indicators. In the second model, both the CWS indicators and the %CWS indicators loaded onto a CBM factor that was distinct from the productivity and quality measures. For the analyses, response type indicators were not included in either model because of previously-discussed problems with collinearity between response type and ideas. The fit of both models was unacceptably poor. Modifications suggested in the modification indices were added to both models until each approached the minimum reasonable fit statistics. However, this required the addition of many correlated error terms for both models. The mediocre fit and high number of correlated errors in this model suggest that additional research may be needed to answer this question satisfactorily.

The results of these analyses may tentatively suggest support for the two-factor model for several reasons. When evaluating the parsimony of the models for the fourth research question, the two-factor model is preferable to the three-factor model. It has fewer dimensions, and it required two fewer correlated error terms to achieve mediocre fit. Lastly, the two-factor model is slightly more interpretable than the three-factor model, both because it has fewer correlated error terms and because it gives inherent meaning to the CBM scores. As an additional consideration, the three-factor model required several correlated error terms between CBM indicators and the indicators from other factors. The fact that CBM indicators may share additional variance with indicators from other factors suggests that they may fit better when modeled as loading onto those other factors. Model fit statistics are presented in Table 4, and the final two-factor

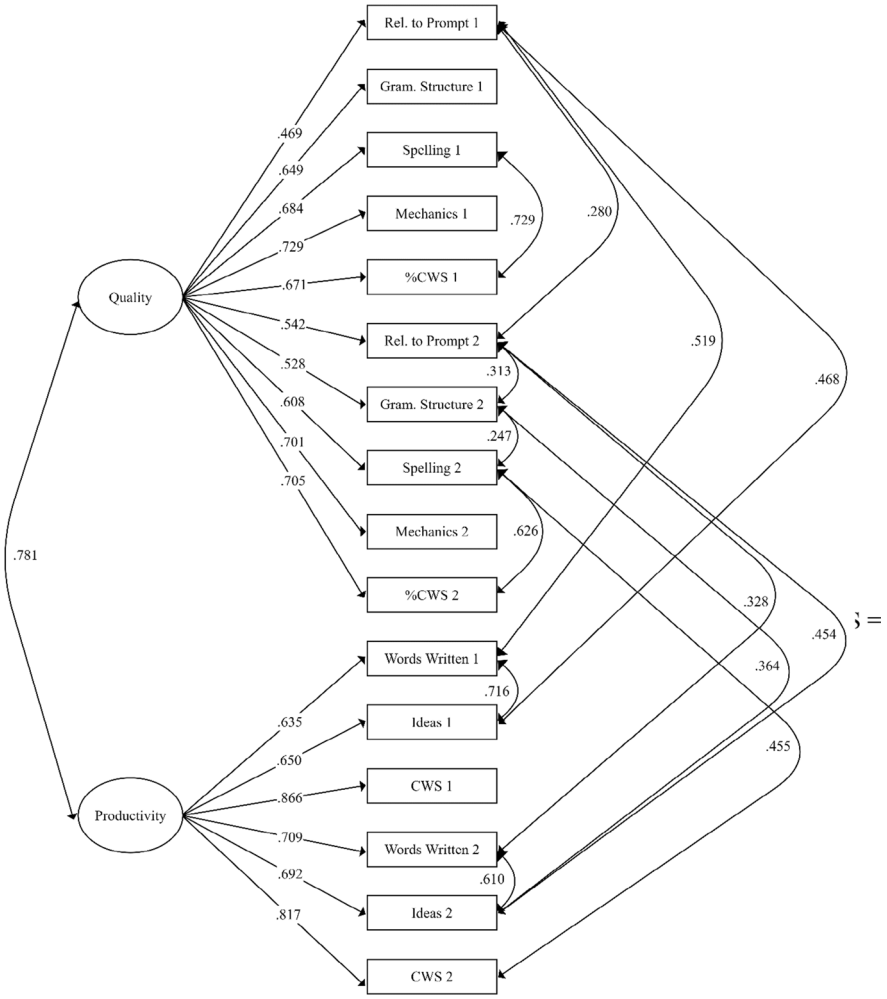

Fig. 3 Standardized factor loadings of the final two-factor model that includes indicators from all three scoring systems. *Rel to Prompt* relationship to prompt, *Gram. Structure* grammatical structure, %CWS percent correct word sequences, *CWS* (number of) correct word sequences

model is presented in Fig. 3: $\chi^2(90) = 331.231$, CFI = .869, TLI = .825, RMSEA = .097 (.086, .109). The results of this model should be interpreted with extreme caution, because the fact that so many parameters needed to be added to the model to achieve even a mediocre fit suggests that the theoretical model may have been a poor starting point for modeling the data.

Discussion

In addition to expectations regarding writing letters, spelling words, and writing sentences, *Common Core State Standards* (Common Core Standards, 2010) also contain expectations for composing text. Kindergarten children are expected to use a range of compositional methods including drawing, dictating, and writing to narrate a single event or several linked events with some details about events and a sense of closure. They are expected to compose informative text and opinion pieces in which they introduce the topic they are writing about and state an opinion about the topic (CCSS, 2010). Yet, we do not have reliable and valid measures for scoring compositions in young, beginning writers in kindergarten. This study contributes the literature on the assessment of composition ability of young, beginning writers, because to date there is little research on the assessment of composition in kindergarten children. This study helps to clarify the dimensionality of a promising qualitative scoring system for compositions that could be particularly useful to teachers because it is quick to administer. Additionally, this study replicates the finding that writing quality and writing productivity are closely related but are nevertheless distinct measures.

This study attempted to replicate Kim and colleagues' (2014) finding with first graders that a qualitative scoring system comprises two distinct dimensions. Despite the fact that the new scoring system used in this study measures similar constructs to the 6 + 1 Traits Rubric that was used in Kim's study, the new scoring system was best modeled as unidimensional. Our findings suggest that the five aspects of the adapted qualitative scoring system cohere to capture a single dimension of substantive quality in the sample of kindergarten children. The dimension captures young children's ability to generate ideas, respond appropriately to the prompt, and use appropriate grammatical structures and transcription skills such as spelling and mechanics. Coker and Ritchey (2010) also reported similar results regarding undimensionality when using a quality rubric to measure sentence writing and concluded that the qualitative score 'assesses multiple proficiencies of beginning writing using a single indicator' (p. 189).

Given the similarities between the present qualitative scoring system and the scoring system used in Kim et al.'s (2014) study, it seems possible that there are substantial differences between the composition abilities of kindergartners and those of the first graders in Kim's study. The high correlation between the accuracy and macro-organization factors in the present study indicates that measures of these two factors covary to such a high degree that they cannot be separated; children with high accuracy almost always have high macro-organization, and children with poor accuracy almost always have poor macro-organization. This could be related to the fact that most of the children in this study wrote short compositions, rarely longer than a couple of sentences. Children who wrote short compositions had few chances to demonstrate technical accuracy. For example, a child who did not write a complete sentence would have received a low score on all three of the macro-organization indicators; although the child may have been able to achieve a high spelling score, he or she would have been unable to score

above one point for mechanics, because a score of two or higher requires the correct use of punctuation, which is almost always a sentence-ending period. Few children used commas or other punctuation in their composition.

An alternative but less likely explanation is that this closer link between technical accuracy and macro-organization may be an artifact of the particular scoring system that was used in the present study. Extant research has clearly indicated that the quality of young children's writing is significantly constrained by their transcription skills such as spelling and handwriting (Graham et al., 1997, Puranik & Al Otaiba, 2012). In other words, macro-organization is largely constrained by technical accuracy. Consequently, a more plausible explanation of our results of unidimensionality is that these results may indicate a stronger constraining influence of transcription skills for the kindergarten children in this sample compared to the first graders in Kim's sample (see Berninger, Mizokawa, & Bragg, 1991, for an explanation of the developmental constraints hypothesis in composition). For kindergarteners, quality is so constrained by transcription that it is indistinguishable.

In line with previous research, this study demonstrated that the dimensions of productivity and quality of writing comprise two distinct but correlated dimensions even for young, beginning writers. As Kim, Al Otaiba, Wanzek, et al. (2015) argue, there is a conceptual link between productivity and quality in writing. There is a certain amount of text (productivity) that is required in order to fully convey an idea (quality), and the more text that is included in a composition, the more opportunity there is to expand on ideas and organize them well. However, some students may be relatively verbose writers without necessarily adding to the quality of their piece. In this sample, this was sometimes the case with students who wrote a great deal about a topic unrelated to the prompt, or who wrote about multiple topics that were unrelated to each other. Thus, productivity and quality are both conceptually related and conceptually distinct.

Interestingly, when indicators from the productivity scoring system were included in the model with indicators from the qualitative scoring system, one of the categories from the qualitative scoring system (i.e. the response type category) was correlated with one of the productivity indicators (i.e. the ideas indicator) to such a degree that it had to be dropped from the model to prevent model estimation problems. This indicates that the particular qualitative scoring system used in this study may also have measured some aspects of writing productivity. These characteristics may make the qualitative scoring system particularly useful for educators who want to quickly and easily get a big-picture view of a child's composition ability. The scoring system may also be useful for progress monitoring or placing children in ability groupings, as is the suggested use of CBM. However, unlike CBM, this scoring system can capture aspects of the content of children's writing, such as how closely related the composition is to a prompt. Furthermore, the scoring system categories are more inherently meaningful than CBM indicators. If a teacher sees that a child's compositions consistently receive a low score in a particular category, the teacher can plan instruction about (for example) including additional details in writing. Conversely, when a child's composition consistently receives low CBM scores, it is impossible to tell from the CBM score exactly which aspects of writing should be targeted.

When CBM indicators were added to the previous models, the models could only achieve mediocre fit with the addition of many correlated error terms. Given the poor fit of the data and the inability to use a direct statistical comparison between these models, it is difficult to draw conclusions about which model has better fit. However, the results of this study tentatively suggest support for the two-factor model over the three-factor model for several reasons. The first reason is that the three-factor model required a high number of cross-dimension factors in order to achieve mediocre fit, including many correlated errors between the CBM indicators and the indicators for productivity and quality dimensions. This may suggest that the CBM indicators share too much variance with the indicators from other factors to be modeled separately. The second reason that the two-factor model may be preferable is that it gives more meaning to the CBM indicators. If a scoring system measures something about writing that is distinct from the components that researchers and educators consider important (such as quality), it is less useful than a scoring system that measures a meaningful component. Considering CBM scores as indicators of meaningful components of writing, such as quality and productivity, rather than considering them as indicators of nothing more than an overall CBM score, assigns the indicators meaning and makes the model more easily interpretable. Of course, due to the relatively poor fit of the models from this paper, future research is necessary to determine how well these CBM indicators actually measure the meaningful components of writing (if at all). Choosing a more interpretable model is not useful if the model does not actually represent the data well.

Previous researchers have questioned the reliability of CBM for young writers (e.g. McMaster & Espin, 2007), despite its prevalence. This questionable reliability may have been one source of the trouble with model fits in this study, particularly since CFA depends on having reliable measures (Kline, 2016). The present study attempted to control for error of measurement by including two essays and several measures of each construct, but these attempts were apparently not sufficient for improving an already error-prone measure. The reliability of CBM for writing is highly dependent on the number of words students generate (Jewell & Malcki, 2005). This may have been the other source of the trouble since the kindergarten children in our sample wrote few words. Coker and Ritchey (2010) also reported this issue in their study with kindergarten children even when children were only producing sentences. Criterion validity for CBM measures for sentence writing ranged from .20 to .30, and were lower that the criterion validity for the Quality and productivity scores. Thus, whereas CBM measures may be appropriate for measuring letter writing, sound spelling, spelling, and sentence writing in kindergarten children it appears to be less appropriate for measuring the composition abilities of kindergarten writers.

Indeed, all of the models in this study had relatively low values for model fit indices, and in most cases, even the final models achieved only mediocre fit (Browne & Cudeck, 1993; Hu & Bentler, 1999). The final fit was lower than the minimum values that have been recommended by other experts (e.g. Nye & Drasgow, 2011; Yu, 2002). If the more conservative cut-offs for fit indices had been pursued, the models would have included many parameters that were not specified a priori, and this could risk capitalizing on chance associations present in this

particular sample but not necessarily representative of the population (Brown, 2015). Conversely, an approach that was more conservative with adding parameters to the models would have resulted in rejecting each model outright, providing little additional information for future researchers. Instead, this paper sought to strike a balance between finding a model that was empirically supported by the present data and finding a model that was similar to the models supported by previous studies.

These challenges reflect the difficulty of assessing writing in general, and they underscore the difficulty of assessing writing in young, emerging writers who have limited to modest writing abilities. It also highlights the complexity of assessing writing; useful writing assessments in one grade may not be useful in another grade. This has been demonstrated in this research, in which CBM scores that have been modeled acceptably in other primary grades (Kim, Al Otaiba, Wanzek, et al., 2015; Kim et al., 2018) could not be acceptably fit to the same model for kindergartners. Indeed, the fact that a single method for scoring writing cannot be used in all grades has been shown by other researchers. For example, Jewell and Malecki (2005) found that certain CBM scores were strong predictors of qualitative measures of writing for second-grade students, but not for fourth- and sixth-grade students. Similarly, Parker, Tindal, and Hasbrouck (1991) found that certain CBM scores were suitable as screening measures for struggling writers in fourth grade, but not in the second and third grades. Taken together, the results of these studies indicate that what we know about writers in one grade may not apply to writers in another grade. Accordingly, what we know about first graders, who are also young, beginning writers, does not apply to kindergarten students.

It is clear that we need continued research to determine the best approach to measure composition skills in young children. These assessments might need to be grade specific to capture developmental competencies. The popularity of holistic/analytic rubrics stems mainly from their convenience and general reliability of scoring; however, they are not perfect. Whereas holistic scoring rubric such as the quality scoring rubric used in this study may be efficient and convenient for scoring writing, they also appear to be designed for the majority (i.e., the average student) and may be less sensitive to writing features displayed by an above average student. Let's take an example of a child who attempts to express a complex thought. To do so, he/she may attempt to spell a difficult word which in turn could lead to more spelling errors and a lower score on the spelling dimension compared to the child who used more simple words but spelled them correctly. Perhaps adding a category to measure word choice as is used in the 6 + 1 rubrics, may be useful. Therefore, one very important avenue for future research is the creation of measures that are developmentally sensitive to features displayed by good, average, and poor writers.

Another important avenue for future research is designing a quality rubric that is more sensitive to textual features, a point raised previously by other researchers (e.g., Huot, 2002). Holistic scoring systems have an identical range of scores allocated across the various categories measured. In the quality rubric used in this study, scores across all five categories examined were rated on a score of 0–3. However, some categories might be better scored on a larger range of scores based on developmental writing expectations.

Until such a time, based on the results of the present study including the internal consistency reliability and findings of other research on writing in kindergarten children (e.g., Coker & Ritchey, 2010; Ritchey, 2008), a qualitative and productivity scoring system appear to be reasonably sufficient for measuring kindergarten composition writing. These measures are easy and quick to administer and score, which is an important consideration for school-based research and in-classroom assessment. It allows teachers to make judgements about overall writing quality based on a number of dimensions; dimensions that previous researchers have been found to be important to evaluate. Consequently, a teacher could direct instructional attention to these important aspects of writing (productivity and quality, including accuracy and macro-organization) that a student might be struggling with. A look at our data indicates that the majority of the children obtained a score of zero on the 'relationship to prompt' criteria. Per CCSS standards, kindergarten children are expected to introduce the topic they are writing about and state an opinion about the topic (CCSS, 2010). Clearly our data indicate that many students (approximately 46%) are unable to do so. This may be a dimension of writing that kindergarten teachers may need to specifically focus on during writing instruction.

Finally, because young children's ability to produce written text is severely constrained by their transcription skills, perhaps eliciting ideas orally may reveal organizational capacities that are obscured by tasks that require the production of text. If students exhibit difficulties with generating ideas and organizing thoughts, instruction could focus on these two elements without the additional burden of writing. Once students are able to generate ideas and organize text, teachers could further support the writing process by helping students spell words or forming letters.

Limitations and Directions for Future Research

This study has raised several interesting questions about kindergartner's composition ability that cannot be fully explored with the present data. For example, collecting three or more compositions from children would have allowed for method effects to be included in the model. Including these method effects may have allowed clearer conclusions to be drawn about dimensionality because fewer correlated error terms would have been required. For example, given the inherent, theoretical link between quality and productivity (as well as links between other measures), it may have been beneficial to assume that the quality indicators and productivity indicators of a particular essay would be related, over and above the relation between the quality indicators from multiple essays by the same participant. Being able to model this relationship with a multi-trait, multi-method model may have significantly improved model fit. However, these types of models require either more than two measurements or stringent assumptions about the structure of the data (Brown, 2015; Widaman, 1985) that may have been unmerited in this case.

The discrepancies between the present paper and previous findings were unexpected given the wealth of research supporting similar factors structures for the compositions of slightly older writers (e.g. Hall-Mills & Apel, 2015; Kim et al., 2014; Kim, Al Otaiba, Wanzek, et al., 2015; Wagner et al., 2011). There are several possible

explanations for these differences. The first is that in young children, measurements of complex skills like composition may be inherently error-prone. This problem can sometimes be circumvented by taking several measurements within a short time span so that additional measures of each indicator type can be included in the model. A second possibility is that the composition skills of kindergartners are qualitatively different, so any model of kindergarten composition that is based on models of older elementary school students' composition will be a poor fit. Future researchers may benefit from taking these considerations into account when planning studies.

Children were given a prompt and a short span of time to write about the prompt. This means that any conclusions drawn about the dimensionality of composition ability may only apply to children's ability to compose spontaneously over a short time frame. This is one of the most common methods of measuring composition ability for young children (e.g. Abbott & Berninger, 1993; Graham, Harris, & Fink, 2000; Kent et al., 2014; Kim, Al Otaiba, Wanzek, et al., 2015; Wagner et al., 2011), probably because it may give the purest picture of a child's independent ability. Including scores from a child's compositions for school assignments may provide an interesting supplement to future research.

Finally, the demographic make-up of the sample of children who participated in this study must be acknowledged. The sample was predominantly White. They were typically developing children who came from predominantly English-speaking homes with higher than average family income. Future research should attempt to include a more diverse group of children (e.g., children with writing disabilities, English language learners) to improve external validity of results.

In conclusion, this study contributes additional knowledge in the field of writing assessment by revealing two dimensions of children's written composition for kindergarten students: quality and productivity. It reinforces the usefulness of both qualitative and quantitative scoring systems for compositions for certain educational and research purposes, and it raises questions about the usefulness of CBM in certain situations. In an era when children are expected to read and write at increasingly younger ages, correct understanding of the measurement of writing is imperative.

Acknowledgements This research was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160253 awarded to University of Cincinnati (PI: Guo). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Portions of this work were conducted as part of the second author's doctoral dissertation at Georgia State University.

Appendix

See Table 5.

Table 5 Studies examining dimensionality of young students written composition

Study	Partici-pants' grade level	Statistical method used	Factors explored	Indicators
Hall-Mills and Apel (2015)	2nd–4th	EFA	Productivity	Total words, total T-units, number of different words
			Grammatical complexity	Mean length of T-unit, clauses per sentence, clause density
			Grammatical accuracy	Percentage of grammatical sentences, grammatical errors per T-unit
			Lexical diversity	Lexical density (proportion of content words to total words)
			Macrostructure	Organization, text structure, Cohesion
Kim et al. (2014)	1st	CFA	Quality	Ideas, organization, word choice, sentence fluency
			Spelling and writing conventions	Spelling, mechanics (capitalization, punctuation), handwriting neatness
			Productivity	Number of words, number of different words, number of ideas
			Syntactic complexity	Mean length of T-unit, clause density
Kim, Al Otaiba, Wanzek, et al. (2015)	2nd–3rd	CFA	Quality	Idea quality, organization
			CBM	%CWS, CIWS
			Productivity	Number of words, number of ideas
Puranik et al. (2008)	3rd–6th	EFA	Productivity	Number of words, number of ideas, number of T-units, number of clauses
			Complexity	Mean length of T-unit, clause density
			Accuracy	Proportion of grammatical T-units, proportion of spelling errors, percentage of correct punctuation
Tindal and Parker (1989)	6th–8th	EFA	Productivity	Number of words, number of legible words, words spelled correctly, CWS
			Production-independent	Mean length of continuous CWS, percentage of legible words, percent of words spelled correctly, %CWS
Wagner et al. (2011)	1st and 4th	CFA	Macro-organization	Topic sentence, logical ordering of ideas, number of key elements
			Complexity	Mean length of T-unit, clause density

Table 5 (continued)

Study	Partici-pants' grade level	Statistical method used	Factors explored	Indicators
			Productivity	Number of words, number of different words
			Spelling and punctuation	Number of spelling errors, Number of capitalization errors, Number of errors involving a period
Current study	KG	CFA	Macro-organization	Response type, relationship to prompt, grammatical structure
			Accuracy	Spelling, mechanics
			Productivity	Number of words, Number of ideas
			CBM	CWS, %CWS

EFA exploratory factor analysis, *CFA* confirmatory factor analysis, *CWS* correct word sequences, *IWS* incorrect word sequences, %CWS percent correct word sequences, *CIWS* correct minus incorrect word sequences, *CBM* curriculum based measures, *T-unit* terminable unit

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. https://doi.org/10.1037/0022-0663.85.3.478.

ACT, Inc. (2018). *ACT college & career readiness standards: Writing*. Retrived May 2, 2020 from http://www.act.org/content/dam/act/unsecured/documents/CCRS-WritingStandards.pdf.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 21,* 102–116. https://doi.org/10.1080/10705511.2014.859510.

Berninger, V. W., Fuller, F., & Whitaker, D. (1996). A process model of writing development across the life span. *Educational Psychology Review, 8,* 193–218. https://doi.org/10.1007/BF01464073.

Berninger, V. W., Mizokawa, D. T., & Bragg, R. (1991). Theory-based diagnosis and remediation of writing disabilities. *Journal of School Psychology, 29,* 57–79. https://doi.org/10.1016/0022-4405(91)90016-K.

Berninger, V. W., Yates, C. M., Cartwright, A. C., Rutberg, J., Remy, E., & Abbott, R. D. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing: An Interdisciplinary Journal, 4,* 257–280. https://doi.org/10.1007/BF01027151.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: The Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Thousand Oaks: SAGE Publications.

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. https://doi.org/10.1177/001440291007600203.

Common Core Standards. (2010). Retreived from http://www.corestandards.org/.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37,* 184–192. https://doi.org/10.1177/00224669030370030801.

Dockrell, J., Ricketts, J., Charman, T., & Lindsay, G. (2014). Exploring writing products in students with language impairments and autism spectrum disorders. *Learning and Instruction*, *32*, 8190. https://doi.org/10.1016/j.learninstruc.2014.01.008.

Education Northwest. (2017). *6 + 1 Trait® rubrics*. Retrived May 2, 2020 from http://educationnorthwest.org/traits/traits-rubrics.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 439–492). Charlotte: Information Age Publishing.

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.

Georgia Department of Education. (2015). *Kindergarten english language arts Georgia standards of excellence (ELA GSE)*. Retrived May 2, 2020 from https://www.georgiastandards.org/Georgia-Standards/Frameworks/ELA-Kindergarten-Standards.pdf.

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. https://doi.org/10.1037/0022-0663.89.1.170.

Graham, S., Harris, K. R., & Fink, B. (2000). Is handwriting causally related to learning to write? Treatment of handwriting problems in beginning writers. *Journal of Educational Psychology, 92,* 620–633. https://doi.org/10.1037//0022-0663.92.4.620.

Hall-Mills, S. (2010). *Linguistic feature development in elementary writing: Analysis of microstructure and macrostructure features in a narrative and an expository genre.* Unpublished doctoral dissertation, Florida State University, Tallahassee.

Hall-Mills, S., & Apel, K. (2015). Linguistic feature development across grades and genre in elementary writing. *Language, Speech, and Hearing Services in Schools, 46,* 242–255. https://doi.org/10.1044/2015_LSHSS-14-0043.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2007). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). New York: The Guilford Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria for new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 1–55. https://doi.org/10.1080/10705519909540118.

Jewell, J., & Malcki, 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.

Kent, S., Wanzek, J., Petscher, Y., Al Otaiba, S., & Kim, Y.-S. G. (2014). Writing fluency and quality in kindergarten and first grade: The role of attention, reading, transcription, and oral language. *Reading and Writing: An Interdisciplinary Journal, 27,* 1163–1188. https://doi.org/10.1007/s11145-013-9480-1.

Kim, Y.-S. G., Al Otaiba, S., Folsom, J. S., Greulich, L., & Puranik, C. S. (2014). Evaluating the dimensionality of first-grade written composition. *Journal of Speech, Language, and Hearing Research, 57,* 199–211. https://doi.org/10.1044/1092-4388(2013/12-0152).

Kim, Y.-S. G., Al Otaiba, S., Puranik, C. S., Folsom, J. S., Greulich, L., & Wagner, R. K. (2011). Componential skills of beginning writing: An exploratory study. *Learning and Individual Differences, 21,* 517–525. https://doi.org/10.1016/j.lindif.2011.06.004.

Kim, Y.-S. G., Al Otaiba, S., & Wanzek, J. (2015). Kindergarten predictors of third grade writing. *Learning and Individual Differences, 37,* 27–37. https://doi.org/10.1016/j.lindif/2014.11.009.

Kim, Y.-S. G., Al Otaiba, S., Wanzek, J., & Gatlin, B. (2015). Toward an understanding of dimensions, predictors, and the gender in written composition. *Journal of Educational Psychology, 107,* 79–95. https://doi.org/10.1037/a0037210.

Kim, Y.-S. G., Gatlin, B., Al Otaiba, S., & Wanzek, J. (2018). Theorization and an empirical investigation of the component-based and developmental text writing fluency construct. *Journal of Learning Disabilities, 51,* 320–335. https://doi.org/10.1177/0022219417712016.

Kim, Y.-S. G., Park, C., & Park, Y. (2013). Is academic language use a separate dimension in beginning writing? Evidence from Korean. *Learning and Individual Differences, 27,* 8–15. https://doi.org/10.1016/j.lindif.2013.06.002.

Kim, Y.-S. G., Park, C., & Park, Y. (2015). Dimensions of discourse level oral language skills and their relation to reading comprehension and written composition: An exploratory study. *Reading and Writing: An Interdisciplinary Journal, 28,* 633–654. https://doi.org/10.1007/s11145-015-9542-7.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: The Guilford Press.

Komsta, L., & Novomestky, F. (2015). *Moments: Moments, cumulants, skewness, kurtosis and related tests*. Retrived May 2, 2020 from https://CRAN.R-project.org/package=moments.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220. https://doi.org/10.1207/s15327906mbr3302_1.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education, 41,* 68–84. https://doi.org/10.1177/00224669070410020301.

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. https://doi.org/10.1177/001440291107700203.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus* (Version 8.1). Retrived May 2, 2020 from https://www.statmodel.com/index.shtml.

Nelson, N. W., & Van Meter, A. M. (2007). Measuring written language ability in narrative samples. *Reading and Writing Quarterly, 23,* 287–309. https://doi.org/10.1080/10573560701277807.

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods, 14,* 548–570. https://doi.org/10.1177/1094428110368562.

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. https://doi.org/10.1007/s11145-007-9062-1.

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

Puranik, C. S., & Al Otaiba, S. (2012). Examining the contribution of handwriting and spelling to written expression in kindergarten children. *Reading and Writing: An Interdisciplinary Journal, 25,* 1523–1546. https://doi.org/10.1007/x11145-011-9331-x.

Puranik, C. S., Al Otaiba, S., Sidler, J. F., & Greulich, L. (2014). Exploring the amount and type of writing instruction during language arts instruction in kindergarten classrooms. *Reading and Writing: An Interdisciplinary Journal, 27,* 213–236. https://doi.org/10.1007/s11145-013-9441-8.

Puranik, C. S., Lombardino, L. J., & Altmann, L. J. P. (2008). Assessing the microstructure of written language using a retelling paradigm. *American Journal of Speech Language Pathology, 17,* 107–120. https://doi.org/10.1044/1058-0360(2008/012).

Ritchey, K. (2008). The building blocks of writing: Learning to write letters and spell words. *Reading and Writing: An Interdisciplinary Journal*, *21*, 27–47. https://doi.org/10.1007/s11145-007-9063-0.

RStudio Team. (2016). *RStudio: Integrated development for R* (Version 1.0.136). Boston, MA: RStudio, Inc. http://www.rstudio.com.

Rugg, G. (2007). *Using statistics: A gentle introduction*. New York: McGraw-Hill House.

Tindal, G., & Parker, R. (1989). Assessment of written expression for students in compensatory and special education programs. *The Journal of Special Education, 23,* 169–183. https://doi.org/10.1177/002246698902300204.

Wagner, R. K., Puranik, C. S., Foorman, B. R., Foster, E., Wilson, L. G., Tschinkel, E., et al. (2011). Modeling the development of written language. *Reading and Writing: An Interdisciplinary Journal, 24,* 203–220. https://doi.org/10.1007/s11145-010-9266-7.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9,* 1–26. https://doi.org/10.1177/014662168500900101.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Doctoral dissertation. Retrived May 2, 2020 from http://www.statmodel.com/download/Yudissertation.pdf.