Efficient Measurement of Writing Knowledge with Forced-Choice Tasks: Preliminary Data Using the Student Knowledge of Writing Test

Gary A. Troia[1], Frank R. Lawrence[1], Julie S. Brehmer[2], Kaitlin Glause[3] & Heather L. Reichmuth[4]

[1]Michigan State University | USA

[2]Saginaw Valley State University | USA

[3]Millikin University | USA

[4]University of Southern Main | USA

Abstract: Much of the research that has examined the writing knowledge of school-age students has relied on interviews to ascertain this information, which is problematic because interviews may underestimate breadth and depth of writing knowledge, require lengthy interactions with participants, and do not permit a direct evaluation of a prescribed array of constituent knowledge elements. For these reasons, our goal in this study is to report the development, piloting, and field testing, using a sample of 335 students from grades 4 and 5, of four alternate versions of a writing knowledge assessment—the Student Knowledge of Writing Test (SKOWT)—that uses forced-choice responses to evaluate students' knowledge of writing processes, genre elements, and linguistic features of written language. All versions of the SKOWT demonstrated adequate internal consistency reliability and construct validity based on exploratory factor analyses following deletion of some items. In addition, there was acceptable predictive criterion validity based on associations of SKOWT scores with subtests from the Test of Written Language-4 and measures of narrative, opinion, and informative essay quality. We discuss how the SKOWT might be used in future research and educational practice.

Keywords: writing knowledge, writing quality, assessment, late elementary

journal of
WRITING RESEARCH

Troia, G., Lawrence, F.R., Brehmer, J.S., Glause, K., & Reichmuth, H.L. (2023). Efficient measurement of writing knowledge with forced-choice tasks: preliminary data using the student knowledge of writing test. *Journal of Writing Research, 15*(2), 395-427. DOI: 10.17239/jowr-2023.15.02.06

Contact: Gary A. Troia, Department of Counseling, Educational Psychology and Special Education, 620 Farm Lane, Michigan State University, East Lansing, MI 48824 | USA – gtroia@msu.edu

1. Introduction

There are at least four different types of knowledge that influence writing performance in school-age children. First is topic knowledge, more of which enables the writer to access, retrieve, and use relevant information and critical details to enhance text content (McCutchen, 1986, 2000). Second is genre knowledge—knowledge about specific discourse structures, both at the micro- and macro-levels, employed to organize textual information in ways familiar to others in a discourse community (Langer, 1986). Micro-level genre knowledge includes information regarding specific kinds of vocabulary and syntactic structures associated with specific genres (Donovan, 2001), such as the use of technical vocabulary terms and nominalization plus non-finite verbs in informational texts versus more general vocabulary and finite verbs in narratives, while macro-genre knowledge includes information regarding rhetorical structures to organize content typical of a particular genre type (Hasan, 1984; Pappas, 1993; Stein & Glenn, 1979). Because discourse organization is attuned to specific writing task demands (e.g., a strict word or page limit) and perceived audience requirements (e.g., a lay versus professional discourse community), genre knowledge also includes knowledge about these facets of the writing environment (Holliway & McCutchen, 2004). Third is linguistic knowledge, which entails knowledge about phonology, morphology, orthography, syntax, and semantics that are relevant for handwriting, spelling, vocabulary use, sentence construction, and the formulation of cohesively linked and coherently meaningful larger segments of text (e.g., Apel et al., 2004; Bourassa et al., 2006; Myhill & Watson, 2014; Olinghouse & Graham, 2009). Fourth is metacognitive knowledge that reflects a writer's beliefs about writing and the cognitive processes used in the act of writing (Lin et al., 2007; Raphael et al., 1989). Metacognitive knowledge includes declarative knowledge of what constitutes good writing, procedural knowledge of writing processes that can be deployed for a given task, and conditional knowledge of strategies appropriate under different writing conditions. Metacognitive knowledge plus genre knowledge are frequently referred to as discourse knowledge in the literature (occasionally, linguistic knowledge also is enveloped by this term).

Graham (2006) noted four assertions for which there should be strong evidence to make the case that knowledge and writing are connected: (1) skilled writers should possess more knowledge than less skilled writers, (2) writing knowledge should accumulate with age and schooling, (3) individual differences in knowledge should predict writing performance, and (4) instruction designed to increase knowledge should improve writing performance. He posited that such evidence would support the model of domain learning advanced by Alexander (1997, 2003), in which growth from acclimation (i.e., initial familiarity with overarching domain features) to competence (i.e., capacity to apply principled reasoning to tackle problems associated with a domain) is driven by three interrelated catalysts—knowledge, motivation, and strategic actions. Such evidence also would accord well with Hayes' (1996, 2011) model of writing that gives prominence to varied types of knowledge and their developmental arc. Graham reviewed the extant research available at the time (also see Graham et al., 2018) and found support for the four assertions, though some had more clear, consistent, and convincing evidence (e.g., individual differences in knowledge predict writing performance) than others (e.g., writing knowledge accumulates over time). Below we highlight some key findings related to each type of writing knowledge.

2. Literature Review

2.1 Topic Knowledge

There is research that demonstrates a strong relationship between topic knowledge and writing. Students who know substantively more about a writing topic are capable of producing qualitatively superior texts than their less knowledgeable peers (e.g., Benton et al., 1995; DeGroff, 1987; Langer, 1984; McCutchen, 1986; Mosenthal et al., 1985). For instance, McCutchen (1986) found that elementary and middle school students (grades 4, 6, and 8) who had extensive knowledge of football wrote longer and more coherent texts that contained a greater proportion of game-related actions than those written by students with limited knowledge of football. Voss et al. (1980) compared fictional accounts of a half inning of a baseball game by undergraduates with low and high knowledge of baseball and found low-knowledge writers generated ideas that were unrelated to game actions (e.g., fan reactions) while high-knowledge writers produced more complete and elaborated descriptions of the game with associated topic-relevant details. Likewise, Benton et al. (1995) observed that ninth graders and undergraduates with greater topic knowledge about baseball generated a greater proportion of topic-relevant ideas in their writing plans, whereas students with more limited topic knowledge generated a greater proportion of topic-irrelevant ideas. Langer (1984) suggests that different degrees of topic knowledge predict success on different kinds of writing tasks—students who are asked to recount facts or elaborate on an assigned topic can rely on relatively unorganized strands of topic knowledge to respond successfully, but when asked to perform an in depth analysis or to defend a position, they will need knowledge that is more integrated and hierarchically organized to be successful. Wijekumar et al. (2019) discovered that initial topic knowledge was a robust predictor of writing quantity and quality even when relevant topic knowledge was introduced using source materials for fifth graders; prior knowledge appeared to help students access, select, and evaluate the information contained in the source materials. A number of other studies have demonstrated an association between topic knowledge and writing performance in grade school students and undergraduates (Albin et al., 1996; Olinghouse & Graham, 2009; Olinghouse et al., 2015).

Although topic knowledge is an important contributor to writing performance and can be measured in a straightforward and expeditious manner, discourse knowledge (again, generally metacognitive and genre knowledge) has been found to be a more robust predictor of writing outcomes than topic knowledge (e.g., McCutchen, 1986; Olinghouse et al., 2015). In a sample of ninth graders and university undergraduate students, Benton et al. (1995) assessed their participants' discourse knowledge, baseball topic knowledge and interest, and their narrative writing performance (content quality, thematic maturity, and syntactic maturity) when asked to write a story about a half-inning of baseball. Topic knowledge was related mostly to planning activity (in other words, idea generation) because this type of knowledge was more strongly correlated with narrative content quality. Discourse knowledge, on the other hand, was related mostly to translation of generated ideas into usable language because this type of knowledge was more strongly correlated with the measures of thematic and syntactic maturity. Thus, discourse knowledge appears to be related to domain-relevant writing proficiency factors whereas topic knowledge appears to be related to task-relevant (e.g., prompt) factors that vary greatly within the domain.

2.2 Genre Knowledge

A number of studies have examined genre knowledge as well as its role in writing. Scholars have observed, using mostly oral retell protocols, that children are aware of at least some genre-specific features from an early age and that this knowledge grows more complete, explicit, and integrated with genre-related linguistic forms and rhetorical functions over time with more exposure and instruction (Berman & Slobin, 1994; Duke, 1999; Hemphill et al., 1994; Pappas, 1993). In writing, students demonstrate similar developmental patterns in their use of genre features (Donovan, 2001; Donovan & Smolkin, 2005; Kamberelis, 1999). For instance, Olinghouse and Graham (2009) interviewed second and fourth graders about what components should be included in a story. While both groups emphasized some vague elements such as "beginning", "middle", and "end", compared to the younger students, the older children more frequently mentioned substantive aspects of writing (i.e., content and processes) rather than production aspects of writing (i.e., mechanics; also see Englert et al., 1988; Fidalgo et al., 2008; Saddler & Graham, 2007). Lin et al. (2007) observed that elementary students in grades 2 through 5 failed to differentiate writing genres of narrative, informative, and persuasive, using similar terms to describe what should be included in each (e.g., tell what happened, add details, think of ideas); in contrast, middle schoolers in grades 6 to 8 exhibited a more complete and refined understanding of story structure as they were able to identify elements such as setting, plot, and resolution and how these are interrelated in a well-developed narrative, though their understanding of the other genres lagged behind (also see Gillespie et al., 2013). Somewhat similarly, Klein and Rose (2010) found that students in grades 5 and 6 displayed incomplete knowledge of persuasive and explanatory genres, identifying only one or two features of each.

Olinghouse and Graham (2009) observed a significant association of knowledge of both narrative elements and production aspects of writing with students' story quality in their study (also see Wen & Coker). Olinghouse et al. (2015) examined whether discourse and topic knowledge about outer space separately predicted inclusion of basic genre elements and the overall quality of fifth graders' narrative, persuasive, and informative compositions once variance attributable to the other type of knowledge, topic interest, spelling proficiency, handwriting fluency, text length, and gender was removed. Both types of knowledge uniquely predicted a significant amount of the variance in composition quality and discourse knowledge uniquely predicted significant variance in the number of genre-specific elements included in each genre of writing. Topic knowledge predicted the inclusion of genre-specific elements only in informational texts.

2.3 Metacognitive and Linguistic Knowledge

In several studies in which interviews were employed to gather information about elementary and middle school students' general writing knowledge, using questions pertaining to the characteristics of good writing, how good writers use varied writing processes, how students themselves approach writing tasks, and why some children have trouble writing, the majority of responses by students reflected both the role of substantive processes involved in writing (e.g., gathering and organizing information, attending to audience needs) and the role of transcription factors, such as handwriting, spelling, punctuation, grammar, and text length, though the substantive aspects of composing tended to be mentioned more often, especially among older students who placed less emphasis on transcription (Barbeiro, 2011; Graham et al., 1993; Kos & Maslowski, 2001; Olinghouse & Graham, 2009; Wong et al., 1989; Wray, 1993). Thus, these studies reveal students possess at least rudimentary metacognitive and linguistic knowledge consistent with growing domain expertise. Using similar methods (but see Englert et al., 1988, who employed vignettes of imaginary students struggling with different aspects of the writing process to solicit solutions), students with writing difficulties, including students with disabilities, have been found to possess overall less knowledge and to emphasize form over substance (Graham et al., 1993; Lin et al., 2007; Saddler & Graham, 2007; Schoonen & de Glopper, 1996). Metacognitive and genre knowledge (i.e., discourse knowledge) have been found to be positively correlated with writing performance across genres and ages (e.g., Englert et al., 1988; Gillespie et al., 2013; Malpique & Veiga-Simão, 2016; Saddler & Graham, 2007) and, in multivariate studies, discourse knowledge is a significant unique predictor of writing (e.g., Fidalgo et al., 2008; Olinghouse & Graham, 2009). Of note, Gillespie et al. (2013) found that fifth graders had different levels of writing knowledge depending on the genre, with narrative writing knowledge being the most well developed, and that metacognitive writing knowledge was predictive of genre knowledge after controlling for gender, writing achievement, and students' emphasis on conventions such as grammar, handwriting, and spelling.

Linguistic knowledge, as well, has been found to be strongly associated with writing performance (e.g., Schoonen et al., 2003). In a study with bilingual (Dutch/English) secondary students in the Netherlands, Schoonen and colleagues (2011) found that measures of vocabulary knowledge (semantics), grammar (primarily morphology), and orthography in each language were significantly positively associated with writing proficiency in each language. Likewise, Trapman et al. (2018), using similar measures of linguistic knowledge for a sample of low achieving Dutch secondary students, observed a positive predictive relationship between levels of vocabulary and grammar knowledge (but not orthographic knowledge) and writing performance levels (sum of primary trait quality scores from three writing tasks representing narrative, argumentative, and informative genres) and, importantly, between growth in grammar knowledge and growth in writing performance from grades 7 to 9. In both of these Dutch language studies, the linguistic knowledge measures employed either completion tasks or multiple-choice tasks adapted from earlier research focused on reading. For their orthographic knowledge completion task, students selected the correct missing letter(s) in a word from several choices. For their grammar knowledge completion task, students filled in sentence word gaps using correct grammatical forms of various word classes to show agreement, aspect, number, or time.

2.4 Methodological Challenges

Much of the research that has examined writing knowledge, except that which has focused on topic knowledge and the few studies on linguistic knowledge conducted in the Netherlands noted above, has relied on interviewing students to ascertain their knowledge. This is problematic for at least three reasons. First, interviews probably underestimate breadth and depth of writing knowledge because children may not possess sufficient metalinguistic competence to express their understandings of writing processes, genres, and conventions: They may not be able to use language effectively to share their knowledge. Conversely, in many studies of student knowledge about writing, verbal responses are categorized and tallied and thus students who elaborate more or provide otherwise lengthier responses may be classified as possessing more writing knowledge, which may be misleading. Olinghouse and colleagues (2015) resolved this methodological issue by using response category proportions in their analyses to control for total number of responses across participants. Nevertheless, the issue of how well students are actually capable of communicating the knowledge they possess remains a problematic concern in this area of research. Second, interviews require lengthy interactions with study participants (e.g., 20-30 minutes per child in the study by Olinghouse & Graham, 2009), recording and transcription of those interactions, and qualitative coding of students' responses. Though digital recording devices combined with relatively automated speech-to-text transcription may alleviate some of the burden, interviews still pose an impediment to assessing student writing knowledge at larger scales within schools, districts, and states, and/or when using repeated measurement to evaluate changes in writing knowledge. Third, interviews do not permit a direct evaluation of a prescribed array of constituent knowledge elements—interviews can prompt students to share their general understandings and assumptions about writing, but cannot readily tap discrete knowledge regarding, for instance, how varied punctuation marks are used to set apart certain dependent clauses, join related independent clauses, denote a series, and mark contracted or conjoined words. Moreover, interview questions cannot easily differentiate correct versus incorrect knowledge. Finally, the multiple limitations of interviews as a means of evaluating school-age students' knowledge may impede research in this area, which might partly explain why relatively so few studies on the relations between writing knowledge and writing performance have been conducted, most prior to the 2010s.

A more efficient assessment format using forced-choice responses to evaluate students' knowledge of writing processes, genre elements, and linguistic features of written language seems to be a reasonable solution to the problems noted above. Such an approach to assessing writing knowledge is not new—many education entities have employed this approach in attempts to quantify students' writing-related knowledge either in combination with direct constructed-response writing performance measures or even in lieu of such performance measures (in the latter case, the objective knowledge test serves as an indirect proxy for actual composing based on evidence that performance on objective measures is usually moderately correlated with quality scores on written papers; e.g., Stiggins, 1982). In a recent review of 49 states' writing assessments, Behizadeh and Pang (2016) found that only three states used forced-choice objective writing knowledge tests in isolation; the remainder used on-demand written essay assessments, but often in combination with writing knowledge tests and/or brief written responses. Thus, assessing writing knowledge continues to be valued by educators and policymakers.

2.5 Research Objectives for This Study

Our goal in this study is to evaluate the reliability and validity of a multiple-choice assessment of writing knowledge for 4th and 5th graders that might be used to measure individual writing-related linguistic, genre, and metacognitive knowledge.

After briefly reporting about item development and pilot testing, we describe our field data collection and statistical analytic procedures to determine the psychometric properties of four alternate forms of the Student Knowledge of Writing Test (SKOWT), and we present information related to the construct validity and predictive criterion validity of the SKOWT. Four versions were developed for a larger research project in which annual growth in writing knowledge (as well as writing motivation and performance) was of interest. We address four research questions in our study. First, does the SKOWT reliably measure the intended knowledge constructs of interest? Second, are scores on the four versions of the SKOWT correlated? Third, is performance on the SKOWT strongly associated with performance on another valid measure of writing knowledge? Fourth, does performance on the SKOWT predict writing performance, specifically writing quality?

3. Method

3.1 Item Development and Pilot Testing

Retired items from 12 different U.S. state writing knowledge tests administered prior to 2010 to students in grades three through six were collected and examined as potential items for the SKOWT. We collected tests used with younger and older students than our intended population of 4th and 5th graders in an attempt to have an acceptable range of item difficulty. To be considered, the items had to use a multiple-choice response format and had to clearly measure knowledge of one of the following six constructs: (1) spelling, (2) capitalization, (3) punctuation, and (4) grammar (i.e., word- and sentence-level linguistic knowledge), (5) genre elements and structures, or (6) writing processes and strategies (i.e., metacognitive knowledge). The first four constructs were considered part of the broad knowledge domain of writing mechanics, whereas the last two constructs were considered part of the broad knowledge domain of discourse. Examples of items used for each of these constructs appear in Figure 1. Items from only a dozen state tests were considered because other states may not have used writing knowledge tests, or did not release items for public use, or were unwilling to share restricted-use test items with the authors. We also created additional items that mimicked the structure and content foci of the items from these state tests to create an adequately large initial pool of potential test items. All items were independently reviewed by three content experts in writing instruction and assessment for item and response choice clarity, appropriateness, and construct relevance. Following this initial round of review, changes to items and response choices were made based on feedback and some items were dropped. The retained items were randomly assigned to one of four versions of the SKOWT, each form containing about 30-32 items total, with typically 4-6 items per assessed construct noted above. Four versions were deemed necessary to provide an opportunity to evaluate student writing knowledge growth across a school year—fitting individual growth curves requires a minimum of four data points.

Although we had initially planned to pilot the entire item pool with a small group of students before developing the four versions, this approach was deemed untenable because students would almost certainly not be willing to take a test with over 120 items; the average testing time for a 30-item test is nearly a half hour. Thus, we elected to use a form administration schedule in which two versions of the SKOWT were administered to small groups of students in grades 4 and 5 recruited from two elementary schools in mid-Michigan (12 students took Forms A and B, 11 students took Forms A and C, 0 students took Forms A and D due to an oversight, 16 students took Forms B and C, 19 students took Forms B and D, and 20 students took Forms C and D) to permit us to rapidly evaluate test-retest alternate form reliability, while one version was randomly administered to another larger group of students from these schools to obtain an initial estimate of internal consistency reliability and to identify poorly performing items to be removed (50 students took Form A, 100 students took Form B, 47 students took Form C, and 0 students took just Form D). Based on the pilot sample of students, the alternate form test-retest reliability using total score was $.41 < r\text{s} < .88$. Although three of the forms' total score correlations were below .60 (Forms B/C, Forms B/D, and Forms A/C), we did not consider this problematic for an initial attempt considering the number of respondents taking two versions. We obtained the following internal consistency reliabilities using all items: $\alpha\text{s} = .92$ for Form A, .81 for Form B, .85 for Form C, and .86 for Form D. Due to weak item-total correlations, we eliminated one writing process item from Form B, one punctuation item from Form C, and one writing process item from Form D. We shifted a few items between forms to arrive at 30 items for each version. All versions of the SKOWT with the correct answers are available at https://osf.io/gu8rd/?view_only=0ef79f5609bb4406b2f832e622317e52.

Table 1. Sample items from the SKOWT

Construct	Capitalization	Punctuation	Spelling	Grammar	Writing Process	Genre
Item stem	Which word in this sentence should be capitalized? In the book Great Explorers, the author discusses the 1926 flight over the North Pole of admiral Richard Byrd.	Which sentence has correct punctuation?	How do you spell the underlined word correctly? Jared put on some <u>descent</u> clothes.	How should this sentence be re-written to correct the grammar mistake? Me and Richard saw two children who looked happy.	Heather wants to write a story about friendship that other people will find interesting to read. To achieve this goal, the first thing Heather does is make a story map (which is a special type of graphic organizer). Heather's story map will be most helpful to her for:	What is the genre of this passage? Katie Oaks is one of the most trusted girls in her class. Maybe they like how Katie listens, or maybe they just feel comfortable around her, but all of the other girls tell their secrets to Katie. When Katie gets a diary for her birthday, she can't help but to write down the secrets that she learns, but when her diary goes missing, Katie finds herself in a tight corner.
Response choices	A) Author B) Book C) Admiral D) Flight	A) Lisa gathered her homework extra pencils and highlighter before going home. B) Lisa gathered her homework, extra pencils, and, highlighter before going home. C) Lisa gathered her homework extra pencils, and highlighter, before going home. D) Lisa gathered her homework, extra pencils, and highlighter before going home.	A) desent B) decent C) dissent D) deecent	A) Richard and me saw two children who looked happy. B) Us and Richard saw two children who looked happy. C) Richard and I saw two children who looked happy. D) They and Richard saw two children who looked happy.	A) Organizing the events in her story. B) Editing her story. C) Picking a title for her story. D) Selecting the best action words to use in her story.	A) Historical fiction B) Science fiction C) Realistic fiction D) Autobiography

Note. Correct response choices are italicized.

3.2 Field Testing

3.2.1 Participants

A total of 397 students from grades 4 (n = 180) and 5 (n = 217) participated in field testing, with ages ranging from 9-0 to 11-2. These students came from 41 general education classrooms distributed throughout 24 different Midwestern U.S. schools. The students in this study were recruited at the classroom level as part of a larger study of the relationships between teachers' writing instructional practices and annual growth in their students' writing performance, knowledge, and motivation. Of the participants, 55% (n = 217) were female and 75% were White (n = 296). Additionally, about 5% (n = 21) of the students were considered non-native English learners and 8% were categorized as students with special needs (n = 31). These students were enrolled in the larger study in four annual cohorts between 2017 and 2022, excluding the 2020-2021 academic year during the COVID-19 pandemic. It is important to note that there are considerable missing data for the SKOWT and writing samples used to evaluate writing performance quality due to (a) student absences, (b) failure to complete testing in several classrooms before the end of the school year, and (c) the beginning of the COVID-19 pandemic in spring of 2020.

3.2.2 Criterion Measures

Writing Performance Quality. Participants' typewritten texts in response to narrative, opinion, and source-referenced informative prompts were hand-scored using a rubric based on the Smarter Balanced Assessment Consortium performance task writing rubrics. The rubric contains seven dimensions: (1) reader orientation to purpose, (2) logical coherence, (3) concluding section, (4) cohesion through linking words or phrases, (5) development of ideas using details such as facts, examples, quotes, and experiences, (6) precise and varied language; and (7) correct grammar/usage/mechanics (i.e., writing conventions). Each dimension was scored on a scale of 0 (no evidence of dimensional quality, severely flawed/incom–prehensible) to 5 (excellent evidence of dimensional quality, virtually no flaws/fully comprehensible) for a total score between 0 and 35. The dimension scores loaded on a single factor that accounted for 57.4%, 58.8%, and 59.0% of total variance for narrative, opinion, and informative papers, respectively. Internal consistency reliabilities using the seven dimensions were .85, .87, and .87 for narrative, opinion, and informative papers, respectively. All the papers were double scored by trained undergraduates and the interrater reliability estimates, calculated with a two-way random effects intraclass correlation using absolute agreement, were .81, .81, and .83 for narrative, opinion, and informative papers, respectively.

Test of Written Language-Fourth Edition (TOWL-4) subtests. The Vocabulary and Spelling & Punctuation subtests of the TOWL-4 were administered to students. For the Vocabulary subtest, students are presented with a word to independently read and then write a single complete sentence using the word exactly as printed (i.e., without altering the tense of part of speech) to demonstrate an understanding of its meaning. For the Spelling & Punctuation subtests (these use the same task but yield separate scores), students transcribe dictated sentences to demonstrate their grasp of written language conventions of spelling/capitalization and punctuation. Raw scores were converted to scaled scores ($M = 10$, $SD = 3$). The Vocabulary subtest has an internal consistency reliability between .85 and .92 for children the same ages as participants in our study, while the Spelling subtest's internal consistency reliability is between .90 and .92 and the Punctuation subtest's is between .91 and .93.

3.2.3 Procedures

All research tasks (writing prompts, TOWL-4 subtests, and SKOWT) were administered to groups of 6-15 students in a quiet room at their local school by a trained graduate research assistant or the first author. The TOWL-4 subtests were administered once at the beginning of the school year, followed within two weeks by the SKOWT and the writing prompts for narrative, opinion, and informative writing samples. The SKOWT form and writing prompt administration for each student was assigned via counterbalancing. Students were asked to respond to the writing prompts on a computer or laptop using a web-based application called Writing Architect (Truckenmiller et al., 2020). Administration of the SKOWT always preceded administration of the writing prompts, and the prompts were delivered over multiple days. Three additional administrations of alternate forms of the SKOWT and writing prompts occurred approximately every two months.

For each writing prompt to which students responded, they were given a printed copy of materials they viewed on the computer screen as well as a blank space below the printed prompt instructions for planning their papers (they were instructed to plan in whatever fashion they had been taught for the genre). Students were permitted up to three minutes to plan each paper and 15 minutes to write. An audible beep paired with a visual warning flashed across the top of the screen was given when one minute remained for the time allotted to writing. All instructions (and passages for informational papers, see below) were not only presented in print and on the computer screen, but also were audibly presented by the computer to help alleviate problems encountered by weaker readers. Students were provided with headphones to listen simultaneously while reading the hard copy and/or electronic versions of materials.

Each genre had four prompt options and students completed all four prompts for each genre by the end of the school year. The prompts (and task instructions noted below) were reviewed by an expert panel of writing researchers and teachers. Narrative prompts were in the form of a story title: (1) One Day of Invisibility; (2) The Attack; (3) Fantastic Voyage; (4) Don't Go into The Attic. Opinion prompts were in the form of a question: (1) Should sugary foods be allowed at school?; (2) Should a person always be honest?; (3) Should cellphones be allowed in classrooms?; (4) Should families be able to pick who their children's friends are? Informative prompts were linked to modified expository passages from online sources. The passage titles were: (1) 13-Year-Old World War II Veteran; (2) Swat Up: Six Reasons to Love Flies; (3) Can an Elevated Bus Solve China's Traffic Woes?; (4) Plastic Bottle Village. Permission was obtained from the copyright holders to use and modify the passages for research. The passages were modified to be within a range of readability appropriate for grades 3 through 8 based on word count, Lexile®, Flesch-Kincaid, and Coh-Metrix degree of narrativity (below 50% for each passage). A pilot study to evaluate the equivalence of these prompts with a sample of approximately 175 children in grades 3 through 8 found no significant differences in text length and quality (including conventions) associated with prompt in any genre.

When responding to a narrative prompt, students were told to "write a creative, fictional story—a make believe story—to match the title; write a story others will find interesting and enjoyable to read and remember, a good story (1) establishes the setting, (2) develops the characters, (3) describes an exciting plot sequence that has a clear beginning event, character actions related to that event, and an outcome or conclusion, and (4) follows the rules of writing." When responding to an opinion prompt, students were told to "write a persuasive essay that convinces readers to agree with your answer to the question and remember, a good persuasive essay (1) clearly states your opinion, (2) gives detailed facts and personal experiences to support your opinion, (3) has a conclusion that helps your readers understand why they should agree with your opinion, and (4) follows the rules of writing." When responding to an informative prompt, students were told to "write an informative paper that will help others learn about the topic of the passage you read; be sure to use information from the article you just read to give reasons why it is important and remember, a well written informative paper (1) has a clear main idea and stays on topic, (2) includes a good introduction and conclusion, (3) uses information from the article stated in your own words plus your own ideas, and (4) follows the rules of writing."

4. Results

Table 1 gives the means and standard deviations for narrative, opinion, and informative writing quality scores, the three TOWL-4 subtest scores, and SKOWT total test scores and scores for each putative broad domain of mechanics (capitalization, punctuation, grammar, and spelling) and discourse (writing genres and processes), as well as *t*-test statistical comparisons of the means for these measures based on grade, gender, non-native English learner status, and special needs status as well as associated effect sizes using Cohen's d. All these variables, based on data from the field testing sample, were normally distributed according to skewness and kurtosis values and inspection of associated histograms and normality plots. Fourth graders displayed significantly lower scores than fifth graders for writing quality in all three genres (ESs ranged from 3.84 to 4.18) and SKOWT mechanics knowledge (ES = 3.64), discourse knowledge (ES = 2.14), and total score (ES = 5.47). Students with special needs performed significantly less well than students without special needs on all subtests of the TOWL (ESs ranged from 2.27 to 2.82), in writing quality across all three genres (ESs between 3.92 and 4.24), and on both broad domains of the SKOWT (ESs = 3.60 for mechanics and 2.12 for discourse) and the total test (ES = 5.40). Girls significantly outperformed boys only in narrative quality (ES = 4.01) and opinion quality (ES = 4.21). Non-native English learners, compared with students whose first language was English, scored significantly lower on the TOWL-4 Vocabulary subtest (ES = 2.82) and in narrative quality (ES = 4.01).

4.1 Does the SKOWT reliably measure the intended knowledge constructs of interest?

Table 2 presents item-level data for each version of the SKOWT as well as summary results from exploratory factor analyses using data from both the pilot and field testing samples (n = 672). Following procedures outlined by Boateng and colleagues (2018), we approached item reduction analysis for the SKOWT, an important step in ascertaining the most parsimonious set of test items associated with the construct(s) of interest, by examining three parameters. First, the corrected item-total (biserial) correlations for each item on each version of the instrument were inspected, and values lower than .250 were flagged (these values are shaded in Table 2). Second, item difficulty and discrimination values, derived using a 2-PL IRT model with maximum likelihood estimation, were inspected. It should be noted that this IRT model is predicated on unidimensionality of the data, an assumption which we subsequently tested (see below).

Using IRT, item difficulty represents the probability of a particular examinee correctly answering a given item, with negative values below .500 indicating greater easiness, positive values above .500 indicating greater difficulty, and values between -.500 and .500 indicating moderate difficulty (Hambleton, 1991). We considered items with absolute difficulty values greater than 1.750 to be potentially too easy or too challenging (these are shaded in Table 2). Item discrimination, on the other hand, represents the slope parameter associated with how steeply the probability of a correct response changes as proficiency increases, with higher values

Table 1. Descriptive statistics and t-test comparisons for key study variables.

	Variable																	
Group	TOWL-4 Vocabulary		TOWL-4 Spelling		TOWL-4 Punctuation		Narrative Quality		Opinion Quality		Informative Quality		SKOWT Mechanics		SKOWT Discourse		SKOWT Total	
Grade	M	SD	M	SD	M	M	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
4	10.34	2.43	9.23	2.86	11.07	11.93	11.35	3.73	12.01	3.78	11.93	4.06	11.35	3.73	6.15	2.13	17.49	5.48
5	10.87	3.14	9.97	3.00	11.15	14.46	13.24	3.57	14.92	4.22	14.46	4.26	13.24	3.57	6.92	2.16	20.16	5.46
t-test	*t*(393.32) = -1.90		*t*(395) = -2.52		*t*(395) = -0.32		*t*(392) = -6.66*		*t*(393) = -7.14*		*t*(390) = -5.98*		*t*(395) = -5.15*		*t*(395) = -3.58*		*t*(395) = -4.83*	
ES (Cohen's *d*)	2.85		2.96		2.35		3.84		4.03		4.18		3.64		2.14		5.47	
Gender	M	SD	M	SD	M	M	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Male	10.78	3.02	9.51	3.13	10.84	12.67	12.09	3.82	12.77	4.04	12.67	4.16	12.09	3.82	6.43	2.27	18.52	5.85
Female	10.50	2.70	9.75	2.81	11.34	13.82	12.62	3.69	14.28	4.36	13.82	4.46	12.62	3.69	6.69	2.09	19.31	5.41
t-test	*t*(395) = 0.98		*t*(395) = -0.76		*t*(395) = -2.13		*t*(392) = -2.97*		*t*(393) = -3.56*		*t*(390) = -2.62		*t*(395) = -1.41		*t*(395) = -1.15		*t*(395) = -1.39	
ES (Cohen's *d*)	2.85		2.96		2.34		4.01		4.21		4.33		3.75		2.17		5.61	
Non-native English Learner	M	SD	M	SD	M	M	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Yes	8.90	2.79	8.57	2.75	10.52	11.89	11.48	3.42	11.50	3.18	11.89	3.92	11.48	3.42	5.33	1.88	16.81	5.15
No	10.73	2.82	9.69	2.96	11.15	13.38	12.43	3.77	13.71	4.30	13.38	4.37	12.43	3.77	6.64	2.17	19.07	5.63
t-test	*t*(395) = -2.88*		*t*(395) = -1.70		*t*(395) = -1.18		*t*(29.71) = -4.55*		*t*(393) = -2.32		*t*(390) = -1.53		*t*(395) = -1.14		*t*(395) = -2.71		*t*(395) = -1.80	
ES (Cohen's *d*)	2.82		2.95		2.35		4.01		4.25		4.35		3.75		2.16		5.61	
Student with Special Needs	M	SD	M	SD	M	M	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Yes	7.87	2.39	6.55	2.64	9.00	9.77	8.63	4.12	9.38	3.77	9.77	4.06	8.63	4.12	4.91	2.57	13.53	6.42
No	10.86	2.76	9.90	2.83	11.29	13.59	12.70	3.55	13.94	4.13	13.59	4.26	12.70	3.55	6.71	2.08	19.41	5.31
t-test	*t*(395) = -5.85*		*t*(395) = -6.35*		*t*(395) = -5.39*		*t*(392) = -5.22*		*t*(393) = -5.85*		*t*(390) = -4.73*		*t*(395) = -6.05*		*t*(395) = -4.54*		*t*(395) = -5.82*	
ES (Cohen's *d*)	2.74		2.82		2.27		3.92		4.11		4.24		3.60		2.12		5.40	

Note. Adjusted *t* values and degrees of freedom reported when equal variances could not be assumed; Bonferroni corrected alpha level was used for each set of group comparisons, with α = 0.006.

indicating greater discrimination. Baker and Kim (2017) note that values less than 0.650 have low discriminatory power, thus we flagged discrimination index values lower than this in Table 2. Items in Table 2 with shading for at least two of these parameters were judged to be poorly functioning and were dropped prior to subsequent analyses.

We next conducted a separate exploratory factor analysis (EFA) of the remaining items on each version of the SKOWT using principal axis factoring (PAF) of the correlation matrix with direct oblimin rotation, as performance on writing-related knowledge factors would be expected to be correlated (see Costello & Osborne, 2005). The assumptions for conducting these EFAs (lack of substantial multicollinearity or singularity, adequate sampling adequacy based on the Kaiser-Meyer-Olkin test, and the observed correlation matrix is not an identity matrix based on Bartlett's Test of Sphericity) were all met. Parallel analysis was employed using the rawpar.sps syntax for SPSS to determine the number of factors to retain in each EFA.

The parallel analysis approach compares eigenvalues generated from the actual data to those obtained via Monte Carlo simulation of an uncorrelated data matrix of the same size and is typically more reliable than using the Kaiser criterion, scree plot inspection, and/or percent of variance explained to judge how many factors should be retained due to tendency to over- or under-extract factors using these methods, though multiple methods are preferred over any single one (Franklin et al., 1995). Specifically, the eigenvalue associated with the 95th percentile generated from the simulation (because parallel analysis tends to retain too many factors when using the mean eigenvalue generated by random data) is compared with the eigenvalue generated from the actual data for each factor; if the observed eigenvalue is greater than that from the simulated data, the factor should be retained because the associated value is significantly greater than the one generated with random data (see Hayton et al., 2004). For SKOWT Forms A and D, the parallel analysis indicated the first two factors should be retained, but for SKOWT Forms B and C, only the first factor had a raw data eigenvalue greater than that at the 95th percentile for simulated data. The scree plots comparing actual versus randomly generated eigenvalues for all four versions of the SKOWT suggested a single factor was sufficient to represent the structure of the data. Additionally, item loadings with two-factor solutions for Forms A and D did not produce interpretable separate factors. Thus, we used a unidimensional factor for each form of the SKOWT; the associated factor loadings are given in Table 2 and those below .300 are shaded; the associated items were removed to rerun the factor analyses to derive the final loading values that also are in Table 2. Finally, the internal consistency reliability for each modified form is noted in this table—Cronbach alphas ranged from .83 to .89.

Table 2. Item descriptive statistics for each form of SKOWT and exploratory factor analysis results.

Item Statistics and Factor Loadings for SKOWT Form A										
Item	n	M	SD	α[1]	Corrected Item Total Correlation[2]	Item Difficulty[3]	Item Discrimination[4]	Initial EFA Factor Loading[5]	Final EFA Factor Loading	Cronbach α
Cap_1	331	.69	.462	.886	.454	-0.771	1.213	.462	.451	
Cap_2	331	.66	.475	.890	.295	-1.158	0.557	.309	.309	
Cap_3	331	.70	.459	.886	.458	-0.879	1.091	.463	.459	
Cap_4	329	.45	.498	.891	.262	0.514	0.596	.257		
Punc_1	330	.60	.491	.889	.321	-0.513	0.725	.318	.323	
Punc_2	330	.83	.376	.887	.457	-1.633	1.179	.497	.498	
Punc_3	329	.73	.446	.885	.535	-0.892	1.359	.546	.541	
Punc_4	328	.65	.479	.887	.444	-0.615	1.035	.468	.460	
Gram_1	328	.70	.460	.890	.297	-1.321	0.647	.304	.298	0.893
Gram_2	328	.78	.412	.885	.551	-1.149	1.491	.597	.596	
Gram_3	324	.71	.453	.884	.550	-0.803	1.445	.581	.580	
Gram_4	322	.67	.471	.890	.272	-1.222	0.579	.290		
Gram_5	323	.76	.427	.885	.518	-1.116	1.287	.548	.544	
Spell_1	304	.63	.484	.887	.448	-0.542	1.038	.466	.464	
Spell_2	302	.74	.438	.889	.323	-1.488	0.767	.338	.347	
Spell_3	302	.59	.493	.891	.231	-0.684	0.496			
Spell_4	302	.69	.464	.888	.381	-1.013	0.850	.386	.382	

Spell_5	302	.74	.442	.884	.567	-0.893	1.592	.595	.599
Genre_1	309	.51	.501	.887	.412	0.050	0.960	.430	.431
Genre_2	305	.73	.444	.885	.412	-0.879	1.588	.569	.572
Genre_3	305	.73	.444	.888	.532	-1.193	0.955	.402	.403
Genre_4	305	.77	.421	.885	.374	-1.118	1.476	.564	.569
Genre_5	303	.40	.490	.890	.520	0.718	0.783	.310	.306
Process_1	323	.74	.441	.885	.302	-0.929	1.402	.594	.601
Process_2	321	.84	.363	.885	.541	-1.297	2.224	.621	.619
Process_3	318	.71	.454	.887	.561	-0.893	1.182	.465	.463
Process_4	315	.70	.458	.885	.434	-0.815	1.260	.553	.559
Process_5	312	.83	.379	.885	.519	-1.347	1.682	.607	.607
Process_6	312	.87	.338	.884	.556	-1.392	2.717	.679	.682
Process_7	311	.80	.398	.884	.607	-1.162	1.869	.641	.644

Table 2. Item descriptive statistics for each form of SKOWT and exploratory factor analysis results.

Item Statistics and Factor Loadings for SKOWT Form B										
Item	n	M	SD	α[1]	Corrected Item Total Correlation[2]	Item Difficulty[3]	Item Discrimination[4]	Initial EFA Factor Loading[5]	Final EFA Factor Loading	Cronbach α
Cap_1	416	.77	.419	.819	.430	-1.199	1.377	.470	.485	0.831
Cap_2	416	.63	.484	.821	.375	-0.649	0.896	.373	.365	
Cap_3	415	.48	.500	.834	.024	1.987	0.037			
Cap_4	415	.89	.314	.824	.287	-2.240	1.139	.348	.408	
Punc_1	412	.41	.492	.826	.234	0.724	0.622			
Punc_2	414	.71	.452	.825	.255	-1.521	0.651	.287		
Punc_3	411	.32	.466	.826	.225	1.383	0.636			
Punc_4	412	.57	.496	.819	.406	-0.292	0.983	.418	.410	
Punc_5	409	.78	.416	.816	.508	-1.210	1.396	.546	.548	
Gram_1	409	.76	.430	.816	.505	-1.103	1.360	.570	.522	
Gram_2	408	.68	.467	.820	.392	-0.998	0.855	.393	.400	
Gram_3	408	.76	.428	.818	.464	-1.269	1.117	.541	.495	
Gram_4	407	.79	.409	.818	.486	-1.100	1.912	.559	.589	
Gram_5	408	.65	.477	.825	.262	-0.953	0.708	.306	.328	
Gram_6	406	.58	.495	.822	.352	-0.366	0.892	.423	.401	
Spell_1	393	.81	.389	.816	.537	-1.400	1.454	.629	.571	

Spell_2	395	.56	.497	.822	.329	-0.402	0.627	.367	.334
Spell_3	394	.78	.415	.818	.486	-1.286	1.289	.561	.537
Spell_4	394	.69	.463	.817	.468	-0.867	1.159	.520	.493
Spell_5	393	.61	.488	.824	.286	-0.712	0.675	.348	.339
Genre_1	398	.68	.468	.818	.456	-0.837	1.071	.494	.442
Genre_2	397	.74	.440	.819	.437	-1.111	1.169	.487	.487
Genre_3	397	.56	.497	.821	.365	-0.262	0.886	.366	.402
Genre_4	393	.52	.500	.825	.269	-0.116	0.570	.279	
Genre_5	394	.70	.460	.819	.436	-0.987	1.005	.464	.458
Process_1	407	.43	.495	.831	.107	0.825	0.406		
Process_2	402	.37	.482	.826	.222	1.400	0.426		
Process_3	402	.49	.500	.826	.228	0.155	0.486		
Process_4	398	.27	.443	.827	.200	1.982	0.563		
Process_5	269	.52	.501	.824	.284	-0.112	0.540	.288	

Table 2. Item descriptive statistics for each form of SKOWT and exploratory factor analysis results.

Item Statistics and Factor Loadings for SKOWT Form C										
Item	n	M	SD	α[1]	Corrected Item Total Correlation[2]	Item Difficulty[3]	Item Discrimination[4]	Initial EFA Factor Loading[5]	Final EFA Factor Loading	Cronbach α
Cap_1	357	.48	.500	.849	.249	0.149	0.534			
Cap_2	358	.70	.460	.844	.409	-1.050	1.004	.407	.401	
Cap_3	359	.66	.475	.848	.260	-0.962	0.800	.304	.298	
Cap_4	358	.92	.264	.844	.453	-1.932	2.404	.472	.472	
Punc_1	359	.86	.347	.844	.412	-1.895	1.293	.456	.453	
Punc_2	358	.61	.487	.844	.396	-0.598	0.974	.409	.402	
Punc_3	357	.75	.435	.841	.492	-1.257	1.122	.517	.515	
Punc_4	355	.49	.501	.848	.284	0.057	0.602	.293		
Punc_5	264	.31	.464	.851	.153	2.151	0.383			0.863
Gram_1	355	.83	.378	.842	.474	-1.715	1.211	.450	.446	
Gram_2	351	.62	.487	.845	.359	-0.731	0.758	.422	.424	
Gram_3	352	.49	.501	.850	.209	-0.087	0.450			
Gram_4	351	.81	.394	.840	.570	-1.268	1.962	.632	.632	
Gram_5	349	.75	.435	.842	.452	-1.258	1.138	.498	.503	
Gram_6	348	.79	.410	.842	.490	-1.396	1.297	.518	.509	
Spell_1	332	.74	.439	.844	.390	-1.259	1.078	.387	.390	

Spell_2	333	.66	.475	.849	.241	-1.542	0.454		
Spell_3	329	.54	.499	.841	.473	-0.190	1.142	.454	.458
Spell_4	331	.78	.415	.842	.481	-1.342	1.298	.525	.521
Spell_5	330	.73	.444	.846	.343	-1.286	0.948	.397	.402
Genre_1	337	.71	.453	.843	.421	-1.120	1.026	.471	.475
Genre_2	338	.55	.498	.844	.396	-0.251	1.110	.465	.464
Genre_3	338	.71	.454	.841	.499	-0.945	1.353	.572	.576
Genre_4	330	.42	.493	.851	.181	0.905	0.392		
Process_1	347	.74	.440	.843	.437	-1.288	1.015	.461	.456
Process_2	343	.72	.448	.844	.389	-1.153	1.054	.468	.469
Process_3	342	.62	.485	.843	.438	-0.586	1.153	.480	.482
Process_4	342	.57	.496	.842	.447	-0.345	1.074	.471	.471
Process_5	340	.67	.470	.841	.488	-0.760	1.371	.550	.555
Process_6	339	.34	.475	.857	-.046	6.453	0.101		

Item Statistics and Factor Loadings for SKOWT Form D

Item	n	M	SD	α[1]	Corrected Item Total Correlation[2]	Item Difficulty[3]	Item Discrimination[4]	Initial EFA Factor Loading[5]	Final EFA Factor Loading	Cronbach α
Cap_1	303	.77	.422	.855	.412	-1.469	1.000	.445	.405	
Cap_2	306	.65	.477	.856	.403	-0.850	0.878	.424	.407	
Cap_3	305	.53	.500	.855	.424	-0.137	1.205	.447	.418	
Cap_4	304	.37	.484	.857	.344	0.717	0.886	.353	.351	
Punc_1	304	.58	.495	.860	.260	-0.582	0.557	.265		
Punc_2	304	.41	.493	.863	.143	1.114	0.321			
Punc_3	305	.62	.485	.855	.440	-0.590	1.051	.459	.452	
Punc_4	304	.64	.481	.855	.439	-0.747	0.898	.485	.454	
Gram_1	302	.87	.340	.855	.449	-1.748	1.558	.511	.542	0.864
Gram_2	302	.68	.466	.854	.473	-0.824	1.225	.515	.532	
Gram_3	302	.52	.501	.855	.416	-0.082	0.928	.438	.444	
Gram_4	302	.74	.440	.853	.508	-1.044	1.383	.562	.550	
Gram_5	301	.66	.474	.855	.425	-0.782	1.073	.442	.447	
Gram_6	302	.51	.501	.857	.367	-0.065	0.922	.387	.395	
Spell_1	287	.73	.446	.857	.337	-1.523	0.739	.362	.383	
Spell_2	287	.82	.386	.858	.300	-1.816	1.028	.332	.387	
Spell_3	287	.80	.402	.858	.327	-1.961	0.812	.367	.352	

Spell_4	285	.69	.461	.857	.369	-0.953	1.133	.395	.412
Spell_5	285	.63	.483	.857	.347	-0.687	0.994	.367	.389
Genre_1	293	.76	.425	.856	.377	-1.356	1.092	.428	.445
Genre_2	291	.39	.488	.860	.260	0.719	0.708	.262	
Genre_3	292	.74	.440	.853	.492	-1.093	1.306	.526	.526
Genre_4	291	.72	.451	.855	.437	-1.111	1.067	.489	.503
Genre_5	290	.32	.469	.858	.302	1.158	0.706	.309	.301
Process_1	302	.81	.392	.855	.437	-1.565	1.220	.481	.522
Process_2	300	.53	.500	.858	.338	-0.186	0.655	.370	.352
Process_3	300	.73	.446	.851	.587	-0.857	1.864	.661	.672
Process_4	297	.70	.460	.854	.462	-1.065	0.946	.505	.454
Process_5	267	.48	.501	.860	.264	0.124	0.606	.271	
Process_6	294	.81	.393	.855	.454				

[1]Coefficient alpha when item is deleted.

[2]Item-total correlations less than .250 are shaded.

[3]Item difficulty values greater than absolute value of 1.750 are shaded.

[4]Item discrimination values below .650 are shaded.

[5]Factor loadings less than .300 are shaded.

We then employed confirmatory factor analysis (CFA) with robust maximum likelihood estimation to determine the model fit for the observed data for the presumed single latent factor for each modified version of the SKOWT using complete data from both the pilot and field-tested samples of 672 children. Evaluation of model fit was based on the following standards noted in Brown (2014): Comparative Fit Index (CFI > 0.95), Tucker Lewis Index (TLI > 0.95), Root Mean Square Error of Approximation (RMSEA < 0.06), and Standardized Root Mean Square Residual (SRMR < 0.08). The fit statistics for Form A are $X^2(324) = 521.75$, $p < .001$, CFI = 0.897, TLI = 0.889, RMSEA = 0.045, and SRMR = 0.049; for Form B, $X^2(170) = 196.40$, $p = .081$, CFI = 0.976, TLI = 0.973, RMSEA = 0.020, and SRMR = 0.039; for Form C, $X^2(230) = 327.18$, $p < .001$, CFI = 0.925, TLI = 0.918, RMSEA = 0.036, and SRMR = 0.048; and for Form D, $X^2(299) = 465.94$, $p < .001$, CFI = 0.869, TLI = 0.857, RMSEA = 0.045, and SRMR = 0.055. Thus, SKOWT Form B exhibited excellent fit for a single-factor solution (but note no writing process items were retained on this form), Form C exhibited good fit, and Forms A and D exhibited marginal fit (with RMSEA and SRMR values within acceptable limits but CFI and TLI values below the desired range).

4.2 Are scores on the four versions of the SKOWT related?

The means and standard deviations for the total scores obtained by the field-tested sample for the modified SKOWT Forms A through D and the Pearson correlations between total scores for each form are presented in Table 3. Based on the means and standard deviations, the versions of the SKOWT are clearly not parallel (i.e., having the same means and variances), in part because of differences in the number of items across versions, but rather should be considered alternate forms. Correlations between versions ranged between .69 and .80, indicative of acceptable but not strong alternate form reliability, suggesting some caution is warranted in assuming a score on one form is comparable to that which would be obtained by the same individual on a different form.

4.3 Is performance on the SKOWT strongly associated with performance on other valid measures of writing knowledge?

To address this research question, we calculated Pearson correlations between total scores on each form of the SKOWT and each TOWL-4 subtest scaled score obtained from children in the field-tested sample. The scores used for analysis were the simple unweighted sums of correct responses on the items associated with the modified total test. As can be observed in Table 3, all correlations were at least moderate in magnitude and statistically significant (all rs > .46), indicating adequate evidence of convergent construct validity. Larger magnitude correlations would not be anticipated considering the restrictive nature of the selected TOWL-4 tasks that solely measure application of writing-related knowledge to sentence writing.

Table 3. Correlations between SKOWT using simple unweighted sums for each form and TOWL-4 subtest scaled scores.

	TOWL-4 Vocabulary	TOWL-4 Spelling	TOWL-4 Punctuation
Mean (SD)	10.63 (2.85)	9.63 (2.96)	11.11 (2.35)
TOWL-4 Vocabulary	1		
TOWL-4 Spelling	.61	1	
TOWL-4 Punctuation	.50	.74	1
Form A Total Score	.58	.63	.59
Form B Total Score	.46	.60	.53
Form C Total Score	.52	.55	.48
Form D Total Score	.48	.60	.51

	Form A Total Score	Form B Total Score	Form C Total Score	Form D Total Score
Mean (SD)	18.39 (6.51)	13.56 (4.56)	15.80 (5.41)	16.84 (5.89)
TOWL-4 Vocabulary				
TOWL-4 Spelling				
TOWL-4 Punctuation				
Form A Total Score	1			
Form B Total Score	.80	1		
Form C Total Score	.78	.69	1	
Form D Total Score	.75	.71	.74	1

Note. All correlations are significant at $p < .001$.

4.4 Does performance on the SKOWT predict writing quality?

Although we collected writing performance data from the field-tested sample of children at multiple time points corresponding to administrations of the SKOWT throughout the school year, we did not believe it was appropriate to correlate SKOWT factor scores with quality of writing in each genre at a single time point because estimates of true writing ability within a genre are notoriously unreliable when measured using a single prompt, especially non-narrative writing (e.g., Bouwer et al., 2015). Thus, we thought it prudent to average quality scores across the four time points to arrive at a more reliable estimate of children's ability in each genre. This also helped address analytic challenges arising from missing data due to the pandemic, student absences, and incomplete testing at the end of the school year.

Total scores, calculated as the simple unweighted sum of correct responses on all items from the modified test version, on Forms A, B, C, and D were significantly (at $p < .01$) correlated with mean writing quality scores for narrative papers (rs = .58, .55, .58, and .56, respectively), opinion papers (rs = .58, .56, .59, and .61, respectively), and informative papers (rs = .59, .60, .59, and .62, respectively). Thus, total scores on the SKOWT predicted anywhere between 30% and 34% of variance in narrative quality, 31% and 37% of variance in opinion quality, and 35% and 38% of variance in informative quality.

5. Discussion

Our goal in this study was to share the development and validation of a more efficient assessment of students' writing-related knowledge than has been used in most prior research, which often has relied on interviews. The SKOWT uses a forced-choice response format to evaluate fourth and fifth graders' knowledge of writing processes, genre elements, and linguistic features of written language (spelling, capitalization, punctuation, and grammar), in line with the approach employed by many state education agencies to supplement their on-demand writing essay assessments (Behizadeh & Pang, 2016). The multiple versions of the SKOWT might permit evaluation of individual growth in writing knowledge; however, in this study we did not directly evaluate the SKOWT for its capacity to assess writing knowledge changes over time.

Pilot and field-testing procedures yielded multiple versions of the SKOWT with adequate internal consistency reliability and construct validity based on exploratory and confirmatory factor analyses. Additionally, sufficient criterion validity for the SKOWT was established through significant positive associations with (a) other assessments of writing ability (the Vocabulary and Spelling & Punctuation subtests of the TOWL-4) and (b) the quality of solicited typed essays in the three main genres of narrative, opinion, and informative writing. Importantly, informative quality was evaluated using a task that required reading a source text to respond to the writing prompt. This was fundamentally different from the writing-only prompt task used to evaluate quality in the other genres. Nevertheless, the magnitude of the correlations between informative quality and SKOWT performance were in line with those using narrative and opinion quality. Unfortunately, the SKOWT can only be used to assess the broad domain of writing knowledge and not discrete constructs associated with mechanics or discourse knowledge. Moreover, the SKOWT cannot be assumed to be appropriate for students in grades other than four or five. Additionally, our sample, though diverse with respect to ethnicity and ability, was relatively small and geographically constrained for instrument validation work. It also must be noted that the lower fit indices for Forms A and D of the SKOWT may be the result of correlated residuals, which would suggest additional common variance shared by groups of items from within each of those forms. Finally, the SKOWT does not evaluate one important aspect of writing knowledge that may be of interest to researchers and educators—knowledge about specific writing topics.

The SKOWT may be of interest to literacy researchers who wish to quickly evaluate the writing-related mechanics and discourse knowledge of study participants in work that (a) examines differences in knowledge across populations of fourth and fifth grade students, (b) evaluates changes in writing knowledge associated with instruction, intervention, or general development, (c) attempts to link writing-related knowledge with other written language abilities such as reading and writing performance, motivation, and achievement, or (d) examines connections between other measures of writing knowledge (e.g., interviews) and the SKOWT. We encourage researchers to use the SKOWT as a template for developing reliable and valid writing knowledge measures for other grades and hope that current versions of the SKOWT, available from the first author, will be further refined through iterative research endeavors as is clearly needed. Likewise, we believe the SKOWT may be a useful classroom assessment instrument to identify individual differences in writing mechanics and discourse knowledge and to track changes in knowledge over time and in response to instruction or intervention. The SKOWT may be of particular interest to those who wish to evaluate writing-related knowledge at scale (i.e., across schools, districts, or states), considering it is based on and aligned with extant indirect measures of writing. A number of mechanisms have been identified through which varied kinds of writing knowledge can be enhanced, including: (1) simply reading a variety of texts (Bereiter & Scardamalia, 1984), (2) carefully scrutinizing exemplary models of writing (e.g., Knudson, 1989), (3) providing insight into the reader's struggle using poorly written text exemplars (Holliway & McCutchen, 2004), (4) observing successful writers compose (Couzijn, 1999), and (5) explicitly teaching specific kinds of writing knowledge (Fidalgo et al., 2008; Fitzgerald & Markham, 1987; Fitzgerald & Teasley, 1986; Gambrell & Chasen, 1991; McCutchen et al., 1997; Mosenthal et al., 1985;

Torrance et al., 2007; Wong et al., 1996). Efforts directed at implementing these types of activities in research studies and classrooms might be further evaluated for their efficacy and effectiveness using the SKOWT.

Acknowledgement

This research was supported in part by grant #R305A160049 from the U.S. Department of Education, Institute of Education Sciences, to Michigan State University. Statements do not necessarily reflect the positions or policies of this agency, and no official endorsement by it should be inferred. We wish to thank the teachers and students who participated in this research project as well as the following research assistants who spent many hours scoring the standardized tests, SKOWT, and writing samples: Taylor Arnold, Norma Eppinger, Mikayla Kalik, Amanda Ling, and Nicole Steele.

References

Albin, M. L., Benton, S. L., & Khramtsova, I. (1996). Individual differences in interest and narrative writing. *Contemporary Educational Psychology*, *21*(4), 305-324. https://doi.org/10.1006/ceps.1996.0024

Alexander, P. A. (1997). Stages and phases of domain learning: The dynamics of subject-matter knowledge, strategy knowledge, and motivation. In C. E. Weinstein & B. L. McCombs (Eds.), *Skill, will, and self-regulation.* Erlbaum.

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher*, *32*(8), 10-14. https://doi.org/10.3102/0013189X032008010

Apel, K., Masterson, J., & Hart, P. (2004). Integration of language components in spelling: Instruction that maximizes students' learning. In E. R. Silliman & L. C. Wilkinson (Eds.), *Language and literacy learning in schools* (pp. 292-315). Guilford.

Baker, F. B., & Kim, S. (2017). *The basics of item response theory using R*. Springer.

Barbeiro, L. F. (2011). What happens when I write? Pupils' writing about writing. *Reading and Writing: An Interdisciplinary Journal*, *24,* 813-834. https://doi.org/10.1007/s11145-010-9226-2

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

Bereiter, C., & Scardamalia, M. (1984). Learning about writing from reading. *Written Communication, 1,* 163-88. https://doi.org/10.1177/0741088384001002001

Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: A cross-linguistic developmental study*. Erlbaum.

Benton, S. L., Corkill, S. J., Sharp, J. M., Downy, R., & Khramstova, I. (1995). Knowledge, interest, and narrative writing. *Journal of Educational Psychology, 87,* 66-79. https://doi.org/10.1037/0022-0663.87.1.66

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6,* 149. https://doi.org/10.3389/fpubh.2018.00149

Bourassa, D., Treiman, R., & Kessler, B. (2006). Use of morphology in spelling by children with dyslexia and typically developing children. *Memory and Cognition, 34,* 703-714. https://doi.org/10.3758/BF03193589

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32* (1), 83-100. https://doi.org/10.1177/0265532214542994

Brown, T. (2014). *Confirmatory factor analysis for applied research*. Guildford Press.

Chi, M. T. H. (1978). Knowledge structures and memory development. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 73-96). Erlbaum.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(7), 1-9. https://doi.org/10.7275/jyj1-4868

Couzijn, M. (1999). Learning to write by observation of writing and reading processes: Effects on learning and transfer. *Learning and Instruction*, *9*, 109-142. https://doi.org/10.1016/S0959-4752(98)00040-1

DeGroff, L. C. (1987). The influence of prior knowledge on writing, conferencing, and revising. *The Elementary School Journal, 88,* 105-118. https://doi.org/10.1086/461527

Donovan, C. A. (2001). Children's development and control of written story and informational genres: Insights from one elementary school. *Research in the Teaching of English*, *35,* 394-447. https://www.jstor.org/stable/40171493

Donovan, C. A., & Smolkin, L. B. (2005). *Elementary school children's written scientific explanations: An examination of developing abilities.* Paper presented at the annual meeting of the National Reading Conference, Miami, FL.

Duke, N. K. (1999). *The scarcity of informational texts in first grade.* (Report No.1-007). Center for the Improvement of Early Reading Achievement. University of Michigan.

Englert, S., Raphael, T., Fear, K., & Anderson, L. (1988). Students' metacognitive knowledge about how to write informational texts. *Learning Disability Quarterly*, *11*, 18-46. https://doi.org/10.2307/1511035

Fidalgo, R., Torrance, M., & Garcia, J. (2008). The long-term effects of writing strategy instruction for young writers. *Contemporary Educational Psychology*, *33*, 672-693. https://doi.org/10.1016/j.cedpsych.2007.09.001

Fitzgerald, J., & Markham, L. R. (1987). Teaching children about revision in writing. *Cognition and Instruction*, *4*, 3-24. https://doi.org/10.1207/s1532690xci0401_1

Fitzgerald, J., & Teasley, A. (1986). Effects of instruction in narrative structure on children's writing. *Journal of Educational Psychology*, *78*, 424-432. https://doi.org/10.1037/0022-0663.78.6.424

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel analysis: A method for determining significant principal components. *Journal of Vegetation Science, 6*(1), 99-106. https://doi.org/10.2307/3236261

Gambrell, L. B., & Chasen, S. P. (1991). Explicit story structure instruction and the narrative writing of fourth- and fifth-grade below-average readers. *Reading Research and Instruction, 31*, 54-62. https://doi.org/10.1080/19388079109558071

Gillespie, A., Olinghouse, N., & Graham, S. (2013). Fifth-grade students' knowledge about writing process and writing genres. *The Elementary School Journal*, *113,* 565-588. https://doi.org/10.1086/669938

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

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., Liu, X., Aitken, A., Ng, C., Bartlett, B., Harris, K. R., & Holzapfel, J. (2018). Effectiveness of literacy programs balancing reading and writing instruction: A meta-analysis. *Reading Research Quarterly*, *53*(3), 279-304. https://doi.org/10.1002/rrq.194

Graham, S., Schwartz, S., & MacArthur, C. A. (1993). Knowledge of writing and the composing process, attitude toward writing, and self-efficacy for students with and without learning disabilities. *Journal of Learning Disabilities*, *26,* 237-249. https://doi.org/10.1177/002221949302600404

Hambleton, R. K. (1991). *Fundamentals of item response theory*. Sage.

Hammill, D. D., & Larsen, S. C. (2009). *Test of written language* (4th ed.). PRO-ED.

Hasan, R. (1984). The nursery tale as a genre. *Nottingham Linguistic Circular, 13,* 71-102. http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=12004084

Hayes, J. R. (1996). A new framework for understanding writing. In C. M. Levy & S. Randall (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp.1-27). Erlbaum.

Hayes, J. R. (2011). Kinds of knowledge-telling: Modeling early writing development. *Journal of Writing Research*, *3*(2), 73-92. https://doi.org/10.17239/jowr-2011.03.02.1

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191-205. https://doi.org/10.1177/1094428104263675

Hemphill, L., Feldman, H. M., Camp, L., Griffin, T., Miranda, A. B., & Wolf, D. P. (1994). Developmental changes in narrative and non-narrative discourse in children with and without brain injury. *Journal of Communication Disorders, 27*, 107-133. https://doi.org/10.1016/0021-9924(94)90037-X

Holliway, D. R., & McCutchen, D. (2004). Audience perspective in young writers' composing and revising. In L. Allal, L. Chanquoy & P. Largy (Eds.), *Revision cognitive and instructional processes* (pp. 87-101). Springer.

Kamberelis, G. (1999). Genre development and learning: Children writing stories, science reports, and poems. *Research in the Teaching of English*, 33, 403-460. https://www.jstor.org/stable/40171463

Klein, P. D., & Rose, M. A. (2010). Teaching argument and explanation to prepare junior students for writing to learn. *Reading Research Quarterly*, *45*(4), 433-461. https://doi.org/10.1598/RRQ.45.4.4

Knudson, R. E. (1989). Effects of instructional strategies on children's informational writing. *Journal of Educational Research*, *83*(2), 91-96. https://doi.org/10.1080/00220671.1989.10885936

Kos, R. & Maslowski, C. (2001). Second graders' perceptions of what is important in writing. *The Elementary School Journal, 101*(5), 567-585. https://doi.org/10.1086/499688

Langer, J. A. (1984). The effects of available information on responses to school writing tasks. *Research in the Teaching of English, 18(1),* 27-44. https://www.jstor.org/stable/40170977

Langer, J. (1986). *Children reading and writing: Structures and strategies.* Ablex.

Lin, S. C., Monroe, B. W., & Troia, G. A. (2007). Development of writing knowledge in grades 2-8: A comparison of typically developing writers and their struggling peers. *Reading and Writing Quarterly, 23,* 207-230. https://doi.org/10.1080/10573560701277542

Malpique, A., & Veiga-Simão, A. M. (2016). Argumentative writing by junior high school students: Discourse knowledge and writing performance. *Infancia y Aprendizaje*, *39*(1), 150-186. https://doi.org/10.1080/02103702.2015.1111609

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language, 25,* 431-444. https://doi.org/10.1016/0749-596X(86)90036-7

McCutchen, D. (2000). Knowledge, processing, and working memory: Implications for a theory of writing. *Educational Psychologist*, *35*(1), 13-23. https://doi.org/10.1207/S15326985EP3501_3

McCutchen, D., Francis, M., & Kerr, S. (1997). Revising for meaning: Effects of knowledge and strategy. *Journal of Educational Psychology*, *89,* 667-676. https://doi.org/10.1037/0022-0663.89.4.667

Monroe, M., & Sherman, E. (1966). *Group diagnostic reading aptitude and achievement tests*. Nevins.

Mosenthal, P., Conley, M., Colella, A., & Davidson-Mosenthal, R. (1985). The influence of prior knowledge and teacher lesson structure on children's production of narratives. *The Elementary School Journal*, *85*, 621-634. https://doi.org/10.1086/461425

Myhill, D. A., & Watson, A. (2014). The role of grammar in the writing curriculum: A review of the literature. *Journal of Child Language Teaching & Therapy, 30*(1), 41-62. https://doi.org/10.1177/0265659013514070

Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology*, *101,* 37-50. https://doi.org/10.1037/a0013462

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

Pappas, C. C. (1993). Is narrative "primary"? Some insights form Kindergarteners' pretend reading of stories and information books. *Journal of Reading Behavior*, *25*, 92-129. https://doi.org/10.1080/10862969309547803

Raphael, T. E., Englert, C. S., & Kirschner, B. W. (1989). Students' metacognitive knowledge about writing. *Research in the Teaching of English*, *23,* 343-379. https://www.jstor.org/stable/40171153

Saddler, B., & Graham, S. (2007). The relationship between writing knowledge and writing performance among more and less skilled writers. *Reading and Writing Quarterly*, *23*(3) 233–249. https://doi.org/10.1080/10573560701277575

Schallert, D. L. (1982). The significance of knowledge: A synthesis of research related to schema theory. In W. Otto & S. Frank (Eds.), *Reading Expository Material* (pp. 13-48). Academic Press.

Schoonen, R., & de Glopper, K. (1996). Writing performance and knowledge about writing. In G. Rijlaarsdam, H. van den Bergh, & M. Couzjin (Eds.), *Theories, models, and methodology in writing research* (pp. 87-106). Amsterdam University Press.

Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning, 53,* 165-202. https://doi.org/10.1111/1467-9922.00213

Schoonen, R., van Gelderen, A., Stoel, R. D., Hulstijn, J., & de Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61,* 31-79. https://doi.org/10.1111/j.1467-9922.2010.00590.x

Stein, N. L., & Glenn, C. C. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *New directions in discourse processing* (Vol.2; pp.53-120). Ablex.

Stiggins, R. J. (1982). A comparison of direct and indirect writing assessment methods. *Research in the Teaching of English, 16*(2), 101-114. https://www.jstor.org/stable/40170937

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford University Press.

Torrance, M., Fidalgo, R., & García, J. N. (2007). The teachability and effectiveness of cognitive self-regulation in sixth-grade writers. *Learning and Instruction*, *17*(3), 265-285. https://doi.org/10.1016/j.learninstruc.2007.02.003

Trapman, M., van Gelderen, A., van Schooten, E., & Hulstijn, J. (2018). Writing proficiency level and writing development of low-achieving adolescents: The roles of linguistic knowledge, fluency, and metacognitive knowledge. *Reading and Writing: An Interdisciplinary Journal, 31,* 893-926. https://doi.org/10.1007/s11145-018-9818-9

Troia, G. A., Harbaugh, A. G., Shankland, R. K., Wolbers, K. A., & Lawrence, A. M. (2013). Relationships between writing motivation, writing activity, and writing performance:

Effects of grade, sex, and ability. *Reading and Writing: An Interdisciplinary Journal*, *26*, 17-44. https://doi.org/10.1007/s11145-012-9379-2

Truckenmiller, A. J., McKindles, J. V., Petscher, Y., Eckert, T. L., & Tock, J. (2020). Expanding curriculum-based measurement in written expression for middle school. *The Journal of Special Education*, *54*(3), 133-145. https://doi.org/10.1177/0022466919887150

Voss, J., Vesonder, G., & Spilich, G. (1980). Text generation and recall by high-knowledge and low-knowledge individuals. *Journal of Verbal Learning and Verbal Behavior, 19,* 651-667. https://doi.org/10.1016/S0022-5371(80)90343-6

Wen, H., & Coker, D. L. (2020). The role of discourse knowledge in writing among first-graders. *Journal of Writing Research, 12*(2), 453-484. https://doi.org/10.17239/jowr-2020.12.02.05

Wijekumar, K., Graham, S., Harris, K. R., Lei, P. W., Barkel, A., Aitken, A. Ray, A., & Houston, J. (2019). The roles of writing knowledge, motivation, strategic behaviors, and skills in predicting elementary students' persuasive writing from source material. *Reading and Writing: An Interdisciplinary Journal*, *32*, 1431-1457. https://doi.org/10.1007/s11145-018-9836-7

Wong, B. Y., Butler, D. L., Ficzere, S. A., & Kuperis, S. (1996). Teaching low achievers and students with learning disabilities to plan, write, and revise opinion essays. *Journal of Learning Disabilities*, *29,* 197-212. https://doi.org/10.1177/002221949602900209

Wong, B., Wong, R., & Blenkinsop, J. (1989). Cognitive and metacognitive aspects of learning disabled adolescents' composing problems. *Learning Disability Quarterly,* 12, 300-322. https://doi.org/10.2307/1510212

Wray, D. (1993). What do children think about writing? *Educational Review, 45*(1), 67-77. https://doi.org/10.1080/0013191930450106