ACCESS for ELLs 2.0 Grades 1–3 Writing Study:

Examining Paper vs. Online Second Language Writing Performance of Grades 1–3 English Language Learners

Prepared by:

Ahyoung Alicia Kim, Ed.D.
Shinhye Lee
Mark Chapman, Ph.D.
Carsten Wilmes, Ph.D.

November 1, 2017

Executive Summary

With enhanced access to technology, large-scale standardized assessments have increasingly transitioned from traditional paper-and-pencil to online platforms. Online tests are delivered via personal computers, laptops, and tablets, and test-takers respond to the test items via these same devices. Online assessments can benefit test administrators and students. The systems allow large numbers of students to take a test at the same time, thereby easing the administrative burden. Online assessments may be designed to leverage technology and deliver tests to students in a way that makes that content more engaging, compared with paper-and-pencil tests.

While online test delivery may offer advantages, some aspects of the technology may disadvantage test-takers, especially for second language writing assessments of English language learners (ELLs), in which students need to type or keyboard extended responses. In the context of ACCESS for ELLs 2.0, the four language domains of listening, speaking, reading, and writing are administered using online delivery and response. In the writing domain only, all Grades 1–3 ELLs take the writing test on paper, while Grade 4–5 students can take the test on paper or online, depending on state or district preference. Although the paper test mode may be optimal for Grades 1–3 ELLs in terms of developmental appropriateness, mixed-mode administrations are operationally complex, require greater planning on behalf of test proctors and developers, and can add considerable cost to the test program. What is essential, and consistent with the WIDA Can Do Philosophy is to gain an understanding of which test mode best allows each student to demonstrate the highest level of writing proficiency. Therefore, the main purpose of this study was to investigate if and how Grades 1–3 ELLs differ in their performance in paper vs. online test modes.

Participants were 149 Grade 1–3 ELLs (Grade 1: $N = 77$; Grade 2: $N = 62$; Grade 3: $N = 10$) from three U.S. schools and five of their educators. The ELLs completed three writing tasks, representing three test modes:

(1) Paper: students completed their writing using a paper-and-pencil format;

(2) Online-Handwriting: students read the prompt online and hand wrote responses; and

(3) Online-Keyboarding: learners read the prompt online and keyboarded responses.

Each 90-minute test administration was limited to 20 ELLs. During each session, researchers observed two or three ELLs and took notes on their writing behaviors. Immediately after the completion of the tasks, researchers interviewed the same students about their preferred test modes. In addition, the five ELL educators who worked with the participating children were interviewed to understand their perception of children's computer familiarity (e.g., typing ability) and educators' preferred test modes. Student performance was analyzed quantitatively (using descriptive statistics and ANOVA) and qualitatively in terms of fluency, accuracy, and complexity. All interview data were analyzed qualitatively.

Findings from quantitative analyses show that ELLs in Grades 1 and 2 received the highest scores under the Paper condition. Specifically, Grade 1 ELLs performed statistically

significantly better on the paper task than on the two online tasks, whereas Grade 2 ELLs performed statistically significantly better on the two handwriting tasks than on the keyboarding task. On the other hand, Grade 3 ELLs performed equally well regardless of the three test modes; however, the sample size of 10 for this grade was relatively small, making it challenging to generalize the findings.

Findings from qualitative analyses yielded somewhat similar patterns, consistent with the quantitative test score results. For instance, regarding fluency, Grades 1–2 children produced the most words on the Paper task and the fewest words on the Online-Keyboarding task. Interview findings provided additional insights. Regardless of the visible difficulties in typing, the majority of ELLs observed (including all Grade 3 ELLs) said they preferred the keyboarding task over the handwriting tasks. These ELLs pointed out the efficiency in writing when keyboarding (e.g., speeded writing), which contradicts the fluency results discussed above. In contrast, the majority of ELL educators preferred the Paper and/or Online-Handwriting modes over the Online-Keyboarding test.

Study results provide practical implications for writing test administration in the context of K–12 ELLs in general and ACCESS for ELLs 2.0 specifically. Findings suggest the paper task is the optimal mode for students in Grade 1. Second-graders performed better when handwriting rather than keyboarding their responses. Insufficient data on students in Grade 3 impedes a concise finding. Overall, for ACCESS for ELLs 2.0, we recommend continuing to administer the writing domain test using the paper test mode to Grades 1–3 students.

Table of Contents

Introduction

With improved technology and enhanced access to technology, large-scale standardized assessments, including second language assessments, have increasingly transitioned from traditional paper-and-pencil to online platforms. Tests are delivered via personal computers, laptops, and tablets, and test-takers respond to the test items via these same devices. Online assessments can have numerous benefits, such as allowing large number of students to take the test at the same time, thereby easing the administrative burden. Computer-delivered assessments may be designed to leverage technology and deliver tests to students in a way that makes their content more engaging, compared with paper-and-pencil tests.

However, online tests may disadvantage test-takers who are not familiar with technology, especially those who are English language learners (ELLs). Second language writing assessments require students to type or keyboard extended responses. Different approaches could be taken to address students' lack of familiarity with computers and typing in online assessments such as ACCESS for ELLs 2.0 (hereafter ACCESS), for example, by having Grades 1–3 students take the writing domain test on paper even though all other domains are administered online; meanwhile, Grades 4–5 students could view the writing prompts online and hand write their responses (online delivery and handwritten response) or keyboard their responses (online delivery and keyboarded response).

The central concern here is the effect of administration and response modes on young children's test performance. That is, does a particular administration and/or response mode have noticeable effects on test outcomes? To date, although a number of studies have been conducted on the topic of paper vs. online writing exams, addressing first language and second language writing tasks, they generally focus on adult learners (e.g., Barkaoui, 2014; Chen, White, McCloskey, Soroui, & Chun, 2011; Laborda, Royo, & Bakieva, 2016; Whithaus, Harrison, & Midyette, 2008). Comparatively few studies exist regarding Grades K–12 ELLs' paper vs. online writing performance (e.g., Choi & Tinkler, 2002; Renn, DeMarco, & MacGregor, 2015), especially in regard to Grades 1–3. Moreover, study findings have been inconsistent regarding the benefits of paper vs. online writing test modes. These studies mostly attended to quantitative measures of performance (e.g., test scores), while leaving much room for speculation as to the actual quality of the students' written test responses (McDonald, 2002).

The main purpose of this study, therefore, was to investigate how ELLs in Grades 1–3 differ in their writing performance on paper vs. online test modes. This study draws on quantitative and qualitative research methods to investigate the mode effect.

Literature Review

Mode Effects in Writing Assessment

Paper-and-pencil language tests have increasingly been replaced with computer-based or online modes. Some advantages of this shift are that online tests can provide immediate feedback and reduce logistical costs associated with administering large-scale assessments (Choi &

Tinkler, 2002; Kröhne & Martens, 2011). What remains underresearched, however, are the mode effects—how different kinds of test administration affect the validity of scores, particularly in the writing section of a language assessment that involves *keyboarding*.

In general, mode effects are based on the idea that "the test administration has a *causal effect on*, for example, estimated competence (i.e., the outcome)" (Kröhne & Martens, 2011, p. 174). Within the context of writing assessment, researchers have drawn on different approaches to investigate this phenomenon, yielding mixed results as to the test scores obtained through each test mode. Some studies found higher ratings on handwritten texts (Burke & Cizek, 2006; Powers, Fowles, Farnum, & Ramsey, 1994); others found higher ratings on keyboarded responses (Lee, 2004; Whithaus et al., 2008; Wolfe, Bolton, Feltovich, & Niday, 1996); and others reported marginal differences between the two modes (Blackhurst, 2005; Breland, Lee, & Muraki, 2004; Green & Maycock, 2004; Weir, O'Sullivan, & Jin, 2007).

At the same time, several studies reported differences in the quality of writing produced through different modes. For instance, writing fluency (as measured by the total number of words and paragraphs in the responses) was greater in the keyboarded texts than the handwritten responses (Russell & Haney, 1997; Russell & Plati, 2001; Wolfe, Bolton, Feltovich, & Niday, 1996). Studies also found *perceived* differences in the quality of writing relative to the media. In some cases, raters held a positive impression of the keyboarded texts for being "cleaner," indicating fewer instances of surface-level errors (spelling) than the handwritten version (Russell & Plati, 1996; Whithaus et al., 2008). However, in other cases, the same surface-level errors were likely to be penalized in the keyboarded texts while being treated more leniently in the handwritten texts (Russell & Tao, 2004). This discrepancy was due to a rating behavior labeled the reader empathy assessment discrepancy effect (Arnold et al., 1990), which stresses "a tendency for readers to identify more personally with the authors of handwritten essays and thus award higher scores" (p. 14). Raters showing such a behavior were often found to relate to handwritten texts at a personal level, positively viewing efforts put into producing them.

Such variability in findings led researchers to explore the relationship between test mode and examinee's computer abilities. Although findings are inconsistent, a consensus of this line of research is that one's computer familiarity, in accordance with computer-related abilities (e.g., keyboarding skills) plays a vital role in test-taking (Horkay, Bennett, Allen, Kaplan, & Yan, 2006; Lee, 2004; Russell & Plati, 2001; Wolfe et al. 1996; Wolfe & Manalo, 2004). For instance, Wolfe et al. (1996) found that with 10th-grade test-takers, mode effects only existed among students with less computer experience—that is, those who were less familiar with keyboarding scored lower on the computer-based test than on the paper-delivered test. They also found that the same-aged test-takers with more keyboarding experiences produced about 50 more words on the computer-based test, while those with fewer experiences wrote 125 more words in the paper task. Russell's (1999) findings corroborated these findings. In the 1999 study, eighth-grade students with slower keyboarding scored lower on the computer-based test and performed better on the paper-based test. With adult language learners, Barkaoui (2014) demonstrated an interaction effect between keyboarding skills and the test-taker's language proficiency. That is,

highly proficient test-takers in English were likely to benefit more from fluent keyboarding skills than those at the lower end of language proficiency. Ling (2017) discovered that the overall efficiency of keyboarding on the TOEFL iBT writing section is affected by adult test-takers' familiarity with a specific type of keyboard. Overall, the study findings suggest that a test-taker's keyboarding skills as well as general computer familiarity affect her or his test performance and/or writing quality.

Young Children and Writing Medium

Extensive research has been conducted on mode effects on writing assessment, yet particularly in the second language testing contexts, less attention has been directed to K–12 examinees and even less to p children in Kindergarten through third grade. Compared to adult test-takers, children in general are likely to have poorer keyboarding skills owing to fewer computer experiences and juvenile motor skills (Donker & Reitsma, 2007). Notably, researchers on first language writing have demonstrated that with young children, handwriting difficulties are likely to transfer to keyboarding proficiency (Freeman, MacKinnon, & Miller, 2005). Bisschop, Morales, Gil, and Jiménez-Suárez (2016) demonstrate a relationship between ability to write by hand and skill with a keyboard: "motor skills are initially less complex on a keyboard than by pen, but when children develop keyboarding skills, the keys are touched in rapid succession, and motor skills become more important" (p. 3). An additional challenge for young ELLs is that they are linguistically developing in their first or second languages (Wolfe & Manalo, 2004). Inevitably, they need to perform "double translation—native language to English to English and then English to keyboard strokes" (Wolfe & Manalo, 2004, p. 55). Another layer for ELLs with different first language backgrounds is that they need to learn the keyboard layout for the second language (Ling, 2017). In this regard, the construct validity of computer-delivered writing assessment for ELLs may be called into question: Do score inferences become more dependent on an ELL's keyboarding proficiency or the child's access to computers in general than on her or his English proficiency (Barkaoui, 2014; Taylor, Jamieson, Eignore, & Kirsch, 1998)? Thus, it becomes critical for ELL testers properly understand how these children perform on computer platforms, and how the medium affects the quality of the written product.

There are few informative studies on first language writing research on early writers (Berninger, Abbott, Augsburger, & Garcia, 2009; Burke & Cizek, 2006; Connelly, Gee, & Walsh, 2007; Crook & Bennett, 2007). These studies attempted to associate the delivery/ transcription mode (computer-keyboarding vs. paper-handwriting) with writing quality (e.g., writing fluency). Connelly et al. (2007) involved 312 schoolchildren ages 4 to 11 years in the United Kingdom to explore how transcribing fluency affects the quality of their essays. This sample of children all received regular computer lessons and keyboarding practice sessions at school. For the study, all children performed both the handwriting and the keyboarding tasks for 2 minutes; the former asked the children to copy a text, while the latter had the children type the same text into the word processor. Writing fluency was examined in terms of the total number of

correct letters produced in each task. Results indicated that regardless of prior keyboarding experiences, children had superior handwriting speed than keyboarding.

Also in the United Kingdom, Crook and Bennett (2007) compared the writing speed of 72 schoolchildren ages 6 to 11 years old in keyboarding and handwriting tasks. In this study, handwriting was observed on a tablet device via a digital pen. Two tasks were administered without a time limit in both transcribing modes. The "no-draft" task had children write two examples of well-practiced text (e.g., writing their own names). For the "draft" task, children were asked to reproduce pangrams (a short sentence using all alphabet letters). Writing speed was measured as the average time children spent to produce texts in each mode. The children were significantly slower in keyboarding than writing with a digitized pen. However, in the simple "no-draft" task, the difference was marginal.

In a study with an accelerated cohort design, Berninger et al. (2009) examined mode effect with a sample of 241 children with and without learning disabilities. The first cohort consisted of children who were in first grade in the beginning of the study and fifth grade at its end. Children in the second cohort first participated as third-graders and exited as seventh-graders. The researchers collected three varieties of writings from the children: alphabet writing, sentence writing, and essay. Data collection took place when children were in second, fourth, and sixth grade. Writing fluency was analyzed through the number of words based on the total time for each task. For sentence and essay writings, syntactic units (e.g., main clauses with independent or dependent clauses) were additionally counted. The finding was that writing fluency differed with regard to task types and grade levels. This was true for sentence writing; children in fourth and sixth grades typed longer sentences while second-graders wrote shorter sentences by keyboarding. Yet for more extensive writing, writing fluency was greater in the handwriting condition than keyboarding for all grade levels. No significant effects were found across the transcribing modes in terms of the number of identified syntactic units.

In the context of second language writing assessment, Renn et al. (2015) is a rare study that investigated the mode effects on elementary-school ELL children. In this small-scale study, 15 ELLs in Grades 4–5 took two writing tasks in keyboarding and handwriting modes each for 20 minutes. In addition to scoring the written responses, the researchers carried out qualitative analysis on the linguistic characteristics (e.g., linguistic stamina, vocabulary use, and language complexity) of the ELLs' writings. In terms of the test scores, raters assessed student writing from the two modes of handwriting vs. keyboarding. In alignment with Connelly et al. (2007), Crook and Bennett (2007), and Berninger et al. (2009), the total number of words and the mean length of utterances were higher for handwritten responses. Notably, error analysis indicated that the keyboarded texts contained more word-level and surface-level errors (e.g., spacing and capitalization errors).

Overall, these four studies suggest that with younger children, the lack of computer abilities may limit the quality of writing as measured through fluency indices. The studies reveal the qualitative differences of the two types of texts, which further illuminates how children actually perform on different composition modes.

Research Purpose and Research Questions

With the exception of a few studies (e.g., Renn et al., 2015; Wilmes, Olsen, & Montee, 2016), young ELLs have rarely been the focus when exploring test mode effects. Yet in-depth investigation is needed on this population of test-takers regarding the prevalence of K–12 computer-mediated writing assessments such as ACCESS and the ongoing demand for test-based accountability of the student group. For ELL test-takers, this call for research is urgent; score interpretations of the ELL-targeted writing tests "would be confounded with ability to use a computer" (Taylor et al., 1998) as these children (relatively more so than adult language learners) lack second language ability *and* computer skills. ELL educators and language testers need to understand whether technological enhancement in large-scale writing assessment constitute a source of construct-irrelevant variance (e.g., heavy dependence on computer familiarity, fluent keyboarding skills) (Taylor et al., 1998). Yet a substantial lack of understanding exists of the test mode effects on young ELLs' writing performance.

The present study, therefore, collects data from multiple sources and conducts triangulated methods of analysis. Often, comparability studies in test modes establish score equivalence across the differing modes, while leaving questions as to qualitative differences between the actual texts produced in their respective modes (McDonald, 2002). In response to this gap in the literature, and to present a comprehensive account of young ELL children's writing performance as a whole, the current study analyzed children's test scores and writing quality in relation to the differing test modes. Participants were examined using three different tasks, combining varying delivery and response modes: (1) Paper task (paper delivery and handwritten response), (2) Online-Handwriting task (online delivery and handwritten response), and (3) Online-Keyboarding task (online delivery and keyboarded response). In addition, key stakeholders' preferences for paper vs. online test modes were explained. The stakeholders included not only ELLs, but also their educators, who are often involved with test administration.

This study addressed these research questions:

1. To what extent do Grade 1–3 ELLs' scores on paper vs. online test modes differ?
2. How do Grade 1–3 ELLs' writing quality on paper vs. online test modes vary in terms of task completion, fluency, accuracy, and complexity?
3. Do Grade 1–3 ELLs display any behavioral differences on paper vs. online test modes?
4. To what extent do Grade 1–3 ELLs prefer paper vs. online test modes?
5. To what extent do Grade 1–3 ELL educators prefer paper vs. online test modes?

Methods

The study involved students and educators from three U.S. schools. The students took three kinds of tests, one on paper, one that mixed handwriting and keyboarding, and one that only involved a computer. Select students were observed. All student writing samples were scored. All five educators and the observed students were interviewed afterward.

Participants

Grades 1–3 ELLs. Participants were Grade 1–3 ELLs from three schools in the U.S. where ACCESS is given as part of state requirements. A total of 149 children participated with varying numbers of children in each grade level (Table 1). There were similar numbers of boys ($N = 76$) and girls ($N = 73$). Home languages varied with the majority speaking Spanish ($N = 137$); others spoke Haitian Creole ($N = 9$), Kanjoval ($N = 2$), and Farsi ($N = 1$).

Table 1. Number of ELLs in Each Grade

Grade	N size
1	77
2	62
3	10

Grades 1–3 ELL educators. Five educators working with Grades 1–3 shared their perceptions of ELLs' writing abilities and educators' preferences for paper vs. online test modes. Three educators were ELL teachers, and one taught English as a second language. The fifth coordinated an ELL program, and supported both ELLs and ELL teachers. As shown in Table 2, the educators' background varied in terms of the grades they supported and their years of teaching experience.

Table 2. Number of ELLs in Each Grade

Educators	Position	Grades Supported	Years of Teaching Experience
Sara	ELL teacher	K–2	6
Terry	ELL teacher	3–8	20
Amy	ESL teacher	K–3, 5	17
Delores	ESL teacher	1–3	30
Valerie	ESL coordinator	K–5	15

Instruments

Writing tasks. To examine ELLs' writing performance across paper vs. online test modes, three comparable writing tasks were adapted from WIDA MODEL[1] Grades 1–2 and 3–5. The tasks were similar in terms of difficulty[2] and format (i.e., all were constructed response tasks requiring students to write extended responses). However, the three tasks varied in terms of topic and test mode as seen in Table 3: (1) Paper task (paper delivery + handwritten response), (2) Online-Handwriting task (online delivery + handwritten response), and (3) Online-Keyboarding

[1] WIDA MODEL is an English language proficiency assessment for K–12 ELLs in the U.S. and abroad. MODEL writing tasks were adapted in this study due to their similarity to ACCESS. Like ACCESS, MODEL measures four domains of speaking, listening, reading, and writing, and the test is available to students in Grades 1–2, 3–5, 6–8, and 9–12. The writing domain of MODEL is similar to ACCESS in terms of difficulty level and format.

[2] Field test data suggest the tasks had similar difficulty. For details, refer to the MODEL field test report (WIDA Consortium, 2012).

task (online delivery + keyboarded response). The paper version was administered using the paper mode, whereas the two online tasks were administered using the online mode. The Online-Handwriting task required the students to view the task on screen and hand write their responses on paper, whereas the Online-Keyboarding task asked them to view the task on screen and type their responses.

Table 3. Writing Tasks

Grade(s)	Paper Task (Paper Delivery + Handwritten Response)	Online-Handwriting Task (Online Delivery + Handwritten Response)	Online-Keyboarding Task (Online Delivery + Keyboarded Response)
1–2	No eggs	Flying kites	Lemonade stand
3	Lion and mouse	Family activities	Berry pancakes

Observation notes. Researchers took observation notes (refer to Appendix A) during test administration to examine behavioral differences that ELLs displayed across the three tasks.

Post-test questions for Grades 1–3 ELLs. Researchers asked ELLs to identify their favorite task and their preference for paper vs. online test modes (refer to Appendix B).

Semi-structured interview questions for Grades 1–3 ELL teachers. Semi-structured interview questions (Appendix C) were used to investigate ELL educators' perceptions of ELLs' writing abilities and educators' preferences for paper vs. online test modes.

Writing scoring rubric. Researchers used the MODEL scoring rubric (Appendix D) to score students' writing performance holistically in the areas of linguistic complexity, vocabulary usage, and language control. *Linguistic complexity* refers to the quantity of the language produced and the organization of the writing. *Vocabulary usage*, as the name indicates, focuses on how well students used their vocabulary to express meaning. Meanwhile, *language control* refers to use of written grammar, word choice, and mechanics. The scores ranged from 1 (entering) to 6 (reaching). Some responses that could not be rated due to illegibility, off-task response, or blank response were indicated as "nonratable." In addition, each score could be rated for strength (+) or weakness (-) in a specific area.

Procedures for Data Collection

ELLs' writing data were collected in Spring 2017. A minimum of two researchers (one an experienced elementary school educator) visited the school site to administer the writing tasks. Each test session took approximately 90 minutes during which ELLs completed all three tasks. The sequence of the tasks administered varied to control for task sequence effect (e.g., some children completed the paper task first whereas others performed one of the online tasks first). The size of each test session was kept to a maximum of 20 ELLs.

During each session, each researcher randomly selected and observed one to two ELLs, and took notes on their writing behaviors on the observation form in Appendix A. They noted, for example, if ELLs were engaged in the tasks or if they displayed varying speed of writing on different tasks. Immediately after the completion of the tasks, researchers asked the observed students about their preferred tasks per the questions in Appendix B: "What was your favorite

activity? Why did you like that one?" "Do you prefer handwriting or typing your activity? Why?" In addition, the five educators who worked with the participating children were interviewed to understand their perceptions of children's computer familiarity (e.g., typing ability) and educators' preference for paper vs. online test modes. Appendix C includes those interview questions.

Procedures for Data Analysis

ELLs' writing samples were scored by trained raters, who referred to the scoring rubric (Appendix D) during the scoring. For quantitative analysis of students' writing, students' scores on a 1–6 rating scale were converted to a 0–18 scale (Table 4) to incorporate students' strengths and weaknesses. Converted scores were further analyzed using SPSS 23. In detail, descriptive statistics were calculated to examine the average mean across the three tasks. Also, a repeated measures ANOVA compared the effect of test mode on ELLs' writing performance.

Table 4. Converted Writing Scale

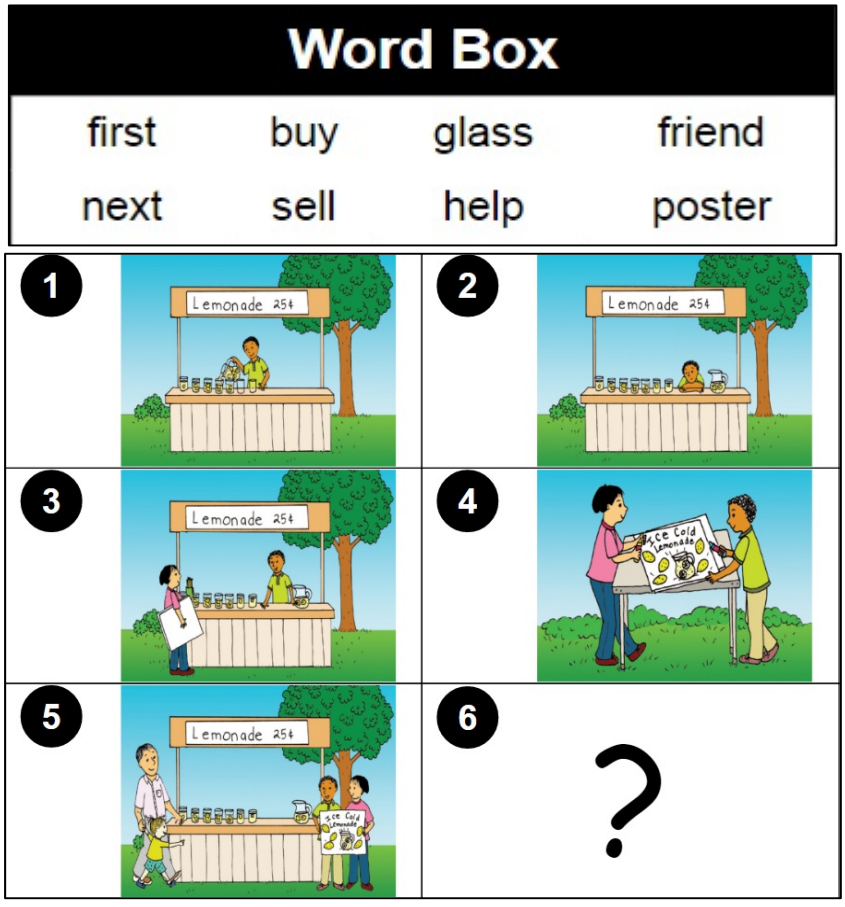
Scale (1-6)	Converted Scale (0-18)
Nonratable	0
1-	1
1	2
1+	3
2-	4
2	5
2+	6
3-	7
3	8
3+	9
4-	10
4	11
4+	12
5-	13
5	14
5+	15
6-	16
6	17
6+	18

To qualitatively analyze students' writing, all handwritten responses from the Paper and Online-Handwriting tasks were transcribed verbatim into word-processing documents by two research assistants. The accuracy of all the transcriptions were double-checked by one of the researchers. Drawing from work on young children's writing (e.g., Bae, 2001; McDonald, 2002), each ELLs' writings were analyzed according to the four measures of *task completion* (the extent ELLs met the requirement of the writing tasks), *fluency* (amount of writing), *accuracy* (the extent

to which ELLs used linguistically appropriate language forms and structures), and *complexity* (syntactic sophistication and lexical diversity).

Task completion was conceptualized as the extent to which ELLs are able to make use of the task input resources in their responses. The picture-description task employed in the present study explicitly required ELLs in Grades 1 and 2 to narrate a story by describing the pictures presented in sequence (i.e., from Picture 1 to Picture 5) while using the provided sample words. For all tasks, the visual information that the ELLs could draw on was eight words presented in a word bank (i.e., Word Box) and five pictures describing each scene as part of a larger story (see Figure 1). Task completion was determined through counting (1) the number of words used from the word bank's eight words and (2) the number of pictures out of the five. Grade 3 ELLs were given a similar task involving pictures, but were not required to describe the pictures; rather they were asked to share an experience that were similar to the pictures.

Figure 1. Word Box and Pictures Presented in the Grades 1 and 2 Test



In terms of *fluency*, the total number of words each ELL produced in each response was counted and then averaged across all ELLs. The counts were lenient to include words displaying orthographic transcription (phonetically influenced by first language orthography or pronunciation; Fashola, Drum, Mayer, & Kang, 1996; Lindgren & Stevenson, 2013) or inventive spellings (e.g., *asc* for *ask*, *naever* for *neighbor*).

A*ccuracy* was broadly defined as two subcategories of errors: grammatical and lexical. *Grammatical errors* were inaccurate usages pertaining to a specific part-of-speech structure (articles, verbs, pronouns, and prepositions) as well as bound morphemes (third-person singular –s, plural –s, possessive 's). The raw frequencies of the errors were counted for each error type, in addition to the total number of errors for the two broader categories of lexical and grammatical errors. With these two units, a ratio of errors per 10 words ([total number of errors / total number of words] * 10) was calculated to normalize the raw error counts of each error type. This method accounted for the impact of text length on skewing the occurrences of errors (a higher word count creates more opportunities for errors; Plakans, Gebril, & Bilki, 2016). In addition, the denominator value of 10 addressed the relatively short length of written responses produced by young ELLs studied (Mehnert, 1998). *Lexical errors* were incorrect uses of words. They were indexed by counting the number of errors that changed the meaning of words. These included misspelled homophones (e.g., *there* for *their*), misspelled words (that are not homophones) that have different meanings from the intended words (e.g., *want* for *went*), awkward word choices (e.g., *put* lemonade in glass), and missing words (e.g., Gave back to ___).

Complexity was analyzed in terms of two components: syntactic complexity and lexical variety. For the former, a T-unit analysis was conducted as described by Hunt (1965). In most cases, a T-unit was identified as one independent clause with its dependent clauses. More specifically, clauses connected with conjunctions (e.g., and, but, or, and so) were broken into T-units. Upon identification, the total number of T-units was tallied for each response, which provided a basis for calculating the corresponding mean T-unit length (total number of T-units / total number of words; Hunt, 1965). Second, mean sentence length (average number of words per sentence) was calculated. Based on Hunt (1965), a sentence boundary was defined as "whatever the student wrote between a capital letter and a period or other end punctuation" (p. 23). These clause- or sentence-level indices were considered to indicate the ELLs' ability to produce longer and broader language units (Richardson, Calnan, Essen, & Lambert, 1976). Third, occurrences of cohesive markers were counted. Cohesive devices refer to the linguistic features linking ideas across clauses and sentences, creating an overall cohesion in the narrative (Bae, 2001; Shapiro & Hudson, 1991). Adapting Bae (2001), the present study explored two types of cohesive devices: conjunctions (e.g., additive, adversative, temporal, causal markers) and references (e.g., pronominal, proper nouns, demonstratives). For *lexical variety*, the unique number of words, type-token ratio ([total number of unique words / total number of words] *100), and the five most frequently used words in each response were identified.

The observation notes taken for ELLs were analyzed to investigate if ELLs demonstrated any noticeable behavioral differences across the three test modes. Data were qualitatively analyzed to identify emerging patterns across the test modes. Data from post-test questions were put into Excel; data were analyzed to understand ELLs' preferred tasks and test modes. Finally, educator interviews were qualitatively analyzed by inputting the data in Excel and categorizing according to the broad themes of (1) computer support ELL received at school, (2) ELLs' computer familiarity, and (3) educators' preferred test mode.

Findings

ELLs' Writing Scores across Tasks

ELLs' writing scores on the three tasks were analyzed using quantitative and qualitative measures. For quantitative analysis, findings from descriptive statistics and repeated measures ANOVA were examined to determine if writing task mode affected ELLs' writing performance.

Descriptive statistics. ELLs' performance across the three tasks varied. Grade 1 and 2 ELLs generally performed best on the paper task compared to the online tasks. Between the two online tasks, the students did better on Online-Handwriting in comparison to Online-Keyboarding. As seen in Table 5, Grade 1 ELLs' mean score on the Paper task ($M = 5.34$) was higher than their Online-Handwriting ($M = 4.42$) or Online-Keyboarding mean scores ($M = 3.96$). In comparison, Grade 2 scores were generally higher than those of Grade 1 ELLs. Similar to Grade 1, Grade 2 ELLs' average on the Paper task ($M = 8.19$) was higher than their Online-Handwriting ($M = 7.56$) or Online-Keyboarding mean scores ($M = 6.42$). However, Grade 3 ELLs performed the best on the Online-Keyboarding task ($M = 8.8$). Their mean on the Paper task ($M = 8.6$) was slightly lower than on the Online-Handwriting ($M = 8.7$) or Online-Keyboarding tasks ($M = 8.8$).

Table 5. Descriptive Statistics of Writing Task Scores

Grade	Task	N	Minimum	Maximum	Mean	Std. Deviation
1	Paper	77	0.00	11.00	5.34	2.90
	Online-Handwriting	77	0.00	11.00	4.42	2.92
	Online-Keyboarding	77	0.00	9.00	3.96	2.76
2	Paper	62	2.00	12.00	8.19	2.49
	Online-Handwriting	62	0.00	11.00	7.56	2.63
	Online-Keyboarding	62	0.00	12.00	6.42	3.29
3	Paper	10	7.00	10.00	8.60	0.84
	Online-Handwriting	10	7.00	11.00	8.70	1.25
	Online-Handwriting	10	7.00	10.00	8.80	0.92

Repeated measures ANOVA. A repeated measures ANOVA was conducted to determine if there were significant differences across the average scores of the three writing test modes. Overall, statistically significant differences were found for Grades 1 and 2 scores, but no differences were found for Grade 3 (see Table 6).

Table 6. Repeated Measures ANOVA – Pairwise Comparisons

Grade	Tasks	Mean Difference	Std. Error	Significance (p)
1	P - OH	0.922[a]	0.237	0.001
	OH - OK	0.455	0.219	0.124
	OK - P	-1.377[a]	0.247	0.000
2	P - OH	0.629	0.217	0.016
	OH - OK	1.145[a]	0.297	0.001
	OK - P	-1.774[a]	0.386	0.000
3	P - OH	-0.100	0.482	1.000
	OH - OK	-0.100	0.379	1.000
	OK - P	0.200	0.442	1.000

[a] The mean difference is significant at $p<.01$ level
Notes: P = Paper task; OH = Online-Handwriting task; OK = Online-Keyboarding task

In detail, for Grade 1 ELLs, a repeated measures ANOVA indicated that the ELLs' writing mean scores differed statistically across test modes ($F(2, 152) = 17.907$, $p < 0.01$). Post hoc tests using the Bonferroni correction revealed that ELLs performed better on the Paper task than on the Online-Handwriting task, which was statistically significant ($p < 0.01$). Similarly, they performed better on the Paper task than on the Online-Keyboarding task, which was statistically significant ($p < 0.01$). Although they performed slightly better on the Online-Handwriting task than on Online-Keyboarding, the score difference was not statistically significant ($p = 0.124$). Therefore, Grade 1 ELLs performed better on paper than on the two online tasks.

Regarding Grade 2 ELLs, a repeated measures ANOVA with a Greenhouse-Geisser correction determined that the ELLs' writing mean scores differed statistically across test modes ($F(1.438, 87.702) = 17.048$, $p < 0.01$). Post hoc tests using the Bonferroni correction revealed that ELLs performed better on the Paper task than on the Online-Keyboarding task, which was statistically significant ($p < 0.01$). Similarly, they performed better on the Online-Handwriting task than on the Online-Keyboarding task, with a statistically significant difference ($p < 0.01$). Although they performed slightly better on the Paper task than on the Online-Keyboarding task, the score difference was not statistically significant at the $p < 0.01$ level. Overall, Grade 2 ELLs performed better on the two handwriting tasks than on the keyboarding task.

For Grade 3 ELLs, a repeated measures ANOVA showed that the ELLs' writing mean scores did not show a statistically significant difference across test modes ($F(2, 18) = 0.105$, $p = 0.901$). In other words, test mode did not affect Grade 3 ELLs' writing performance.

ELLs' Writing Quality

Prior to conducting qualitative analysis, responses were excluded based on five criteria:

1. no response (i.e., left blank);
2. one-word response ("Max");

3. direct copy of test prompts (“max wants to sell lemonade and his friend Susan helps him.Use the pictures towrite a story about max and Susan.”);
4. direct copy of words from the word bank (“next then tree stuck hill fly”); and
5. unintelligible words and/or sentences (“Ibskgtgwgnggsnseubukakgg”).

All Grade 3 ELLs produced responses that could be scored regardless of test mode. This was not the case for Grades 1 and 2 ELLs. More first-graders were excluded than second- or third graders. Table 7 shows the number of excluded responses in Grade 1 and 2 was greater for the two online modes compared to paper. After the exclusions, data were analyzed for 42 Grade 1 ELLs, 50 Grade 2 ELLs, and 10 Grade 3 ELLs.

Table 7. Exclusion Criteria and Example Responses

	Grade 1 (Number of Excluded Responses)			Grade 2 (Number of Excluded Responses)		
Exclusion Criteria	Paper	Online-Handwriting	Online-Keyboarding	Paper	Online-Handwriting	Online-Keyboarding
No response	1	3	2	-	-	4
One-word response	1	2	2	-	-	-
Copy of test prompts	3	5	4	-	-	3
Copy of words from word bank	4	4	-	1	1	1
Unintelligible response	12	15	22	2	3	3
Total	21	29	30	3	4	11

Note: Hyphens indicate no responses fell into specific criteria.

Task completion and fluency. Regarding task completion, Grade 1 and 2 ELLs did not use the majority of word bank words and pictures in their responses. Yet for these ELLs, as seen in Table 8, the average number they used from the word bank’s eight words was the highest in the Paper mode (Grade 1: M = 3.29, Grade 2: M = 4.52) and lowest in the Online-Keyboarding mode (Grade 1: M =2.07; Grade 2: M =3.14). The number of words used also differed by grade level. Grade 2 ELLs applied more of the sample words in their writings than Grade 1 ELLs. Likewise, ELLs described the most pictures in the Paper task (Grade 1: M = 3.88, Grade 2: M = 4.24) and the fewest in the Online-Keyboarding task (Grade 1: M =3.43; Grade 2: M =3.38).

In addition, *fluency* varied across test modes. For Grades 1 and 2, the average number of words was higher on the handwritten responses and the lowest on the keyboarded responses. For example, in both grades, the average number of words was the highest in the Paper task (Grade 1: M = 41.21; Grade 2: M = 53.44), but lowest in the Online-Keyboarding mode (Grade 1: M = 23.98; Grade 2: M = 35.96). On the other hand, Grade 3 ELLs’ length of response did not seem to vary greatly across test modes, while the handwritten responses yielded the highest average number of words (M = 100.20).

Table 8. Descriptive Statistics for Task Completion and Fluency Measures

			Content Coverage (Task Completion)		Fluency
Grade	N	Task	Average Number of Words used from Word Bank (SD)	Average Number of Pictures Described (SD)	Average Number of Words (SD)
1	42	P	3.29 (1.88)	3.88 (1.13)	41.21 (14.42)
	42	OH	2.55 (1.53)	3.76 (1.30)	37.71 (14.20)
	42	OK	2.07 (1.57)	3.43 (1.64)	23.98 (9.71)
2	50	P	4.52 (1.82)	4.24 (0.96)	53.44 (17.40)
	50	OH	3.24 (1.29)	3.96 (1.52)	48.94 (22.32)
	50	OK	3.14 (1.58)	3.38 (1.66)	35.96 (19.00)
3	10	P	-	-	100.20 (25.39)
	10	OH	-	-	89.40 (38.50)
	10	OK	-	-	91.90 (32.57)

Notes: Standard deviations (SD) are in parentheses. The word bank presented eight words. The test included five pictures. P = Paper task; OH = Online-Handwriting task; OK = Online-Keyboarding task

The trends described above are well illustrated in the following excerpts from a Grade 2 ELL's responses. The child produced lengthier texts in the two handwriting modes and provided a relatively detailed description of all pictures in the responses. However, for the keyboarding task, the child not only wrote a shorter response, but partially completed the task by providing incomplete narration of the whole scene (note that the sentences describing a particular picture are marked in numbers of sequential order of the presentation of the pictures in the test).

> [ID: 30063185, Grade 2, Paper Task on "no eggs", score 10, described 5 pictures, 60 words]
> Max get the bowl out for cookies. **(1)** Max go to see the egg there no egg. **(2)** Max and her Mom go to see a person. **(3)** The person gave the egg to Max mom. **(4)** Then Max mom brak the egg and Max stew the egg. **(5)** finally Max and her Mom cooked 1,000 cookies to give everyone an cookies in the world!!!!

> [ID: 30063185, Grade 2, Online-Handwriting Task on "flying kites", score 8, 26 words, described 1 picture]
> when the wind blow 1 kite was stuck **(5)** Then dad got an adia dad got a latter Then dad got it Then the girl was happy

> [ID: 30063185, Grade 2, Online-Keyboarding Task on "lemonade stand", score 8, 29 words, describe 1 picture]
> when the dad and his son buy lemonade **(5)** they drink and give glass then Max say thank you

Accuracy. Results regarding accuracy indicate a mixed trend with regard to the test modes as well as the grade levels. In general, for ELLs in all three grades, the average number of

errors was the highest in the Paper and Online-Handwriting tasks, and the lowest in the Online-Keyboarding task (see Table 9). Yet this pattern is more profound for Grades 2 and 3 ELLs, and especially Grade 3. For Grade 2 ELLs, the average number of *grammatical* errors is similar in the two handwritten texts (Paper: $M = 5.26$; Online-Handwriting: $M = 5.44$), while the keyboarded texts showcased a decrease in the error counts (Online-Keyboarding: $M = 3.68$). In terms of Grade 3, the average number of grammatical error counts is similar in the two online modes (Online-Handwriting: $M = 4.90$; Online-Keyboarding: $M = 4.10$) while the average increased in the Paper task ($M = 7.40$). However, these number of errors could be sensitive to the total length of the writing. Therefore, errors were also counted per 10 words. When interpreting the results from errors per 10 words, the observed differences across test modes become marginal for both lexical and grammatical errors. For all grade levels, *lexical* errors occurred an average of less than once per 10 words in all types of responses. In terms of grammatical errors, Grades 1 and 2 ELLs averaged approximately one error per 10 words in the two online tasks. Tables 10 and 11 provide more detail.

Complexity. Complexity was examined in terms of syntactic complexity (i.e., T-unit, sentence length, cohesive markers, including reference and conjunctive markers) and lexical variety (unique words, type-token ratio). As reported in Table 12, for Grades 1 and 2 ELLs, the *total number of T-units* were generally highest for the Paper task (Grade 1: $M = 6.31$; Grade 2: $M = 7.66$) followed by the Online-Handwriting task (Grade 1: $M = 5.79$; Grade 2: $M = 6.78$), while the lowest was for the Online-Keyboarding task (Grade 1: $M = 4.38$; Grade 2: $M = 5.30$). On the other hand, test mode effects were less evident for *mean length of T-unit* across grade levels. Regardless of test modes, the number of words incorporated per T-unit was consistent. Notably, this measure indicated a clear developmental pattern according to the advancement in grade levels, with Grade 3 ELLs showing the lengthiest T-units in their responses (Grade 1 = 5.68–6.60; Grade 2 = 7.07–7.23; Grade 3 = 8.21–8.37).

Table 9. Descriptive Statistics for Overall Accuracy

			Lexical Errors		Grammatical Errors		Total Errors
Grade	N	Task	Average Number of Errors (SD)	Errors per 10 Words (SD)	Average Number of Errors (SD)	Errors per 10 Words (SD)	Total Errors per 10 Words (SD)
1	42	Paper	1.83 (1.40)	0.47 (0.40)	3.38 (1.95)	0.86 (0.54)	1.26 (0.72)
	42	Online-Handwriting	1.98 (1.63)	0.54 (0.50)	3.90 (2.16)	1.07 (0.60)	1.59 (0.72)
	42	Online-Keyboarding	1.33 (1.24)	0.61 (0.62)	2.79 (2.08)	1.33 (1.43)	1.82 (1.56)
2	50	Paper	2.18 (1.66)	0.41 (0.30)	5.26 (2.46)	1.02 (0.45)	1.39 (0.72)
	50	Online-Handwriting	2.04 (1.51)	0.45 (0.41)	5.44 (2.32)	1.22 (0.54)	1.58 (0.71)
	50	Online-Keyboarding	1.62 (1.32)	0.49 (0.46)	3.68 (2.12)	1.17 (0.64)	1.66 (0.99)
3	10	Paper	3.00 (2.20)	0.47 (0.21)	7.40 (3.43)	0.75 (0.33)	1.06 (0.38)
	10	Online-Handwriting	3.00 (2.83)	0.35 (0.36)	4.90 (4.46)	0.57 (0.46)	0.92 (0.73)
	10	Online-Keyboarding	1.90 (1.45)	0.19 (0.12)	4.10 (2.35)	0.44 (0.19)	0.63 (0.26)

Note: Standard deviations (SD) are in parentheses.

Table 10. Descriptive Statistics for Grammatical Errors

Grade	N	Task	Article (SD)	Verb (SD)	Pronoun, Case, Reference (SD)	Bound Morpheme (SD)	Double Negation (SD)	Preposition (SD)	Average Num. of Errors (SD)	Errors per 10 Words (SD)
1	42	P	0.50 (0.63)	1.52 (1.35)	0.26 (0.50)	0.95 (1.09)	0.02 (0.15)	0.12 (0.40)	3.38 (1.95)	0.86 (0.54)
	42	OH	1.05 (0.88)	1.45 (1.21)	0.12 (0.40)	0.83 (0.89)	0 (0)	0.36 (0.53)	3.90 (2.16)	1.07 (0.60)
	42	OK	0.38 (0.54)	1.10 (1.27)	0.31 (0.56)	0.86 (1.00)	0.02 (0.15)	0.10 (0.30)	2.79 (2.08)	1.33 (1.43)
2	50	P	0.98 (0.89)	1.68 (1.36)	0.52 (0.71)	1.78 (1.15)	0.04 (0.20)	0.26 (0.49)	5.26 (2.46)	1.02 (0.45)
	50	OH	1.60 (1.28)	1.62 (1.19)	0.26 (0.44)	1.46 (1.25)	0.02 (0.14)	0.48 (0.68)	5.44 (2.32)	1.22 (0.54)
	50	OK	0.66 (0.82)	1.38 (1.28)	0.40 (0.78)	1.04 (0.92)	0 (0)	0.20 (0.40)	3.68 (2.12)	1.17 (0.64)
3	10	P	0.60 (0.71)	3.30 (1.64)	0.80 (0.97)	2.00 (1.58)	0.10 (0.33)	0.60 (1.12)	7.40 (3.43)	0.75 (0.33)
	10	OH	0.80 (0.79)	1.60 (2.37)	0.10 (0.32)	1.70 (1.89)	0 (0)	0.70 (1.25)	4.90 (4.46)	0.57 (0.46)
	10	OK	0.20 (0.44)	1.70 (1.33)	0.30 (0.50)	1.50 (1.88)	0.10 (0)	0.30 (0.71)	4.10 (2.35)	0.44 (0.19)

Notes. Standard deviations (SD) are in parentheses. P = Paper task; OH = Online-Handwriting task; OK = Online-Keyboarding task

Table 11. Descriptive Statistics for Lexical Errors

Grade	N	Task	Homophones (SD)	Different Meaning (SD)	Word Choice (SD)	Missing Words (SD)
1	42	Paper	0.40 (0.59)	0.76 (0.82)	0.40 (0.73)	0.26 (0.45)
	42	Online-Handwriting	0.21 (0.47)	1.02 (1.09)	0.29 (0.51)	0.45 (0.63)
	42	Online-Keyboarding	0.31 (0.52)	0.48 (0.74)	0.36 (0.62)	0.19 (0.40)
2	50	Paper	0.40 (0.57)	0.88 (0.92)	0.48 (0.74)	0.42 (0.70)
	50	Online-Handwriting	0.42 (0.50)	1.00 (1.03)	0.26 (0.60)	0.36 (0.69)
	50	Online-Keyboarding	0.30 (0.54)	0.38 (0.57)	0.54 (0.76)	0.40 (0.64)
3	10	Paper	0.60 (1.41)	0.90 (1.12)	0.60 (0.73)	0.90 (1.36)
	10	Online-Handwriting	0.60 (0.84)	1.60 (2.01)	0.40 (0.84)	0.40 (0.70)
	10	Online-Keyboarding	0.20 (0.44)	0.30 (0.50)	0.40 (0.88)	1.00 (1.05)

Note. Standard deviations (SD) are in parentheses.

Table 12. Descriptive Statistics for Syntactic Complexity

Grade	N	Task	Total Number of T-Unit (SD)	Mean Length of T-Unit (SD)	Mean Sentence Length (SD)	Average Number of Reference Markers (SD)	Average Number of Conjunctive Markers (SD)
1	42	Paper	6.31 (1.94)	6.60 (1.52)	30.11 (16.29)	5.40 (2.30)	3.60 (2.12)
	42	Online-Handwriting	5.79 (1.82)	6.54 (1.44)	30.00 (16.83)	3.12 (1.68)	2.22 (2.13)
	42	Online-Keyboarding	4.38 (1.89)	5.68 (2.08)	22.69 (10.08)	2.46 (1.36)	2.24 (2.27)
2	50	Paper	7.66 (2.41)	7.07 (1.16)	27.06 (17.10)	6.92 (2.81)	5.62 (2.97)
	50	Online-Handwriting	6.78 (2.58)	7.23 (1.80)	25.65 (17.21)	3.38 (2.49)	4.66 (3.70)
	50	Online-Keyboarding	5.30 (2.79)	7.07 (2.17)	29.96 (16.14)	4.00 (2.69)	3.78 (3.25)
3	10	Paper	12.20 (3.19)	8.37 (1.38)	33.52 (32.08)	9.40 (4.36)	8.80 (5.66)
	10	Online-Handwriting	11.80 (6.68)	8.24 (1.76)	39.22 (37.72)	8.40 (3.20)	9.20 (5.92)
	10	Online-Keyboarding	11.60 (4.94)	8.21 (1.05)	47.74 (32.94)	6.50 (6.45)	8.60 (5.01)

Note. Standard deviations (SD) are in parentheses.

In terms of *mean sentence length*, a distinct pattern was not observed from any grade level. Yet the data suggest ELLs generally produced sentences each containing more than 20 words across tasks. While the result may indicate that the children wrote fairly long sentences, it should be acknowledged that children have only developing levels of control over punctuation; such a tendency is developmentally appropriate for early writers (Hunt, 1965). In some cases, ELLs' entire writing consisted of one run-on sentence that lacked punctuation marks at sentence boundaries; in other cases, ELLs used multiple conjunctions (e.g., and) rather than periods to distinguish boundaries. The excerpt below showcases such a response from a Grade 3 student, who produced a 45-word sentence:

> [ID: 10063309, Grade 3, Online-Keyboarding task on "lemonade stand", score 6]
> the boy was makeing lemonade store and the boy was wateing the pepone to buy lemonade and nowon wot to bide lemonade and a girl went to hiping the boy and the girl side we neet a peepr because the pepoin and man the gir

Regarding *cohesive* devices, both references and conjunctive markers were examined. The average number of *reference markers* (e.g., proper nouns: Max, Susan) exerted notable differences, especially between the Paper and the Online-Keyboarding tasks. For all grade levels, the number of reference markers was the greatest in the Paper condition (Grade 1: $M = 5.40$; Grade 2: $M = 6.92$; Grade 3: $M = 9.40$). Yet, Grades 1 and 3 ELLs used the lowest number of reference markers in the Online-Keyboarding mode (Grade 1: $M = 2.46$; Grade 3: $M = 6.50$). For *conjunctive markers*, Grade 3 ELLs seemed to use a balanced number in all three tasks, while Grades 1 and 2 ELLs used them more in the Paper condition. This is noticeable given that the word bank across all three tasks for Grades 1 and 2 always contained two to three cohesive devices. For example, in the following two excerpts, the Grade 1 ELL appears to use more conjunctions in the Paper task, whereas such devices were rarely used for the Online-Keyboarding task (note, the originally intended words are marked in brackets).

> [ID: 30063157, Grade 1, Paper task on "no eggs", score 11]
> Max and her mom get the bowl. **next** they dident have eggs. **then** they wak outside. They borrow some eggs from a neighbor. They were douing cookies. **finally** theye bake the cookies **and** They it <eat> the cookies.

> [ID: 60063167, Grade 1, Online-Keyboarding task on "lemonade stand", score 5]
> the boy put the lemonade on the cup.no bodiy bay some lemonade. hise friend come they make a bord forsome bodiy bay a lemonade.**end <and>** some bodiy bay a lemonade.**end <and>** they have moniy

Regarding *lexical variety*, for Grades 1 and 2, the *average number of unique words* was the highest for the Paper task and the lowest for the Online-Keyboarding task (see Table 13). For Grade 3, the average number of unique words were the highest for Paper task ($M = 54.00$), while

the two online tasks were similar in the average number of unique words used. The following two excerpts illustrate the difference in the total number of words as well as the number of unique words used by a Grade 1 ELL across the two modes. For instance, the handwritten text contained a total of 42 words with 28 unique words while the keyboarded text had a total of nine words with nine unique words.

> [ID: 60063167, Grade 1, Paper task on "no eggs", score 10, 42 total words, 28 unique words]
> First Max and his mom were gowing to make cookies put there were no eggs then Max and his mom go to getth eggs ther neyver hod then want home and finally they make the and selepreyt. The cookies with the friends.

> [ID: 60063167, Grade 1, Online-Keyboarding task on "lemonade stand", score 5, 9 total words, 9 unique words]
> max was weyting for some peopel to help hem

Table 13. Lexical Variety in Students' Writing

Grade	N	Task	Average Number of Unique Words (SD)	Type-token Ratio (SD)
1	42	Paper	26.98 (8.56)	67.21 (11.92)
	42	Online-Handwriting	24.15 (9.16)	66.75 (9.65)
	42	Online-Keyboarding	18.30 (7.16)	82.69 (10.71)
2	50	Paper	32.96 (8.84)	63.74 (13.08)
	50	Online-Handwriting	30.34 (11.48)	65.12 (13.61)
	50	Online-Keyboarding	29.96 (16.14)	73.45 (17.60)
3	10	Paper	54.00 (11.07)	56.38 (16.80)
	10	Online-Handwriting	47.60 (13.40)	56.93 (10.88)
	10	Online-Keyboarding	47.90 (11.79)	55.18 (11.27)

Note. Standard deviations (SD) are in parentheses.

The average type-token ratio for Grades 1 and 2 was relatively similar for the two handwriting tasks (Grade 1 Paper: $M = 67.21$, Grade 1 Online-Handwriting: $M = 66.75$; Grade 2 Paper: $M = 63.74$, Grade 2 Online-Handwriting: $M = 65.12$), but much larger for the keyboarding task (Grade 1: $M = 82.69$; Grade 2: $M = 73.45$). This inflation in the proportion of the unique words reflects shorter T-unit and sentence length observed in the Online-Keyboarding task. Generally, the type-token ratio is known to be sensitive to the denominator; that is, the longer the text the more repetition of lexical words (McCarthy & Jarvis, 2010). Grade 3 ELLs' average type-token ratio held consistent regardless of test mode, which reflects the less prominent differences in the average number of words compared to Grades 1 and 2 (see Table 8).

As seen in Table 14, vocabulary frequency indicated the type of words ELLs frequently used in their writings. For each piece of students' writing, the top five most used words were

analyzed. Table 13 indicates that, across test modes, Grades 1 and 2 ELLs were likely to use a great deal of concrete nouns, which are specifically related to the content of the tasks at hand. Grade 3 ELLs used verbs as frequently as concrete nouns across all three tasks. The low frequency of verbs written by Grades 1 and 2 ELLs indicates that Grade 3 ELLs' writings contained greater variety of lexical words, perhaps because they produced lengthier texts for which content was not limited by the input resources (e.g., sample words, task pictures).

Table 14. Vocabulary Frequency

Grade	Task	N	Frequently used words (frequency count)
1	P	42	Eggs (70), Cookies (64), They (37), Mom (29), Max (23)
	OH	42	Kite (76), Fly (21), Got (20), Dad (19), Book (15)
	OK	42	Lemonade (44), Max (21), Buy (17), Friend (15), Came (15)
2	P	50	Eggs (91), Cookies (81), Mom (65), Max (64), Neighbor (29)
	OH	50	Kite (75), Girl (31), Dad (32), Boy (31), Fly (23)
	OK	50	Lemonade (81), Max (43), Buy (31), Came (31), Friend (10)
3	P	10	Help (16), Brother (13), Said (12), Got (10), Day (7)
	OH	10	Went (26), Day (8), Ate (5), Fun (5), Beach (2)
	OK	10	Said (9), Like (8), Ate (8), Mom (7), Eat (6)

P = Paper task; OH = Online-Handwriting task; OK = Online-Keyboarding task

ELLs' Writing Behavior across Tasks

Forty-four ELLs' (Grade 1: $N = 20$, Grade 2: $N = 16$, Grade 3: $N = 8$) writing behaviors were observed while they worked in each test mode. Findings indicated that ELLs in all three grades were quite engaged with the writing activity in the Paper mode. These ELLs seemed to be comfortable performing the paper task, producing lengthier responses than on the two online tasks. On the Online-Handwriting task, Grades 1 and 2 ELLs seemed to spend some time getting familiarized with the onscreen features (e.g., word box, task pictures). Some ELLs appeared to be less engaged or even confused when navigating between screens of the pictures. On the Online-Keyboarding task, these same first and second-graders needed more time because of limited keyboarding skills; 14 out of 36 typed slowly using one finger from one or both hands. They also frequently looked at the keyboard to find appropriate letters. As a result, Grades 1 and 2 ELLs generally produced shorter responses on the keyboarding task than on handwriting tasks. However, Grade 3 ELLs did not demonstrate much behavioral differences on Paper vs. Online-Handwriting tasks.

Table 15 summarizes observations of Grade 1 ELL (ID: 50063467), the student "struggles to find alphabet from keyboard" as indicated in Table 15. Consequently, he produces longer responses on the Paper and the Online-Handwriting tasks than on the Online-Keyboarding task. While his responses are in well-formed paragraphs in the first two tasks, on the keyboarding task, he barely finishes his sentence.

Table 15. Sample Student's Observation Notes and Written Responses

	Paper	Online-Handwriting	Online-Keyboarding
Observation notes	• Student was engaged in the task • Student produces a paragraph	• Student writes a paragraph • Student writes longer compared to keyboard task	• Student tries to copy and paste words from word box • Student struggles to find alphabet from keyboard but writes in sentences • Student accidentally erases some of his writing
Student's response	first max mom has a bowl next they don't hav eggs they went to the neighbor to ask Do you have eggs so we can borrow then they were baking finally they a movie	the boy and the grile are looking in the book then their Dad help them next they fly the kite then one got stuck in a tree then they both share.	first max was selling lemonade on th

ELLs' Preference of Writing Tasks and Mode

After the entire testing session, ELLs were asked about which task they preferred. Grades 1 and 2 ELLs' responses were mixed in terms of their preference for handwriting vs. keyboarding tasks. Approximately two-thirds of the ELLs preferred keyboarding tasks because it was "fun," "it doesn't hurt hands," "it doesn't get messy," and "it is faster to write." However, some expressed difficulty with typing; these ELLs seemed to agree that the difficulty stemmed from locating the letters on the keyboard ("It's hard to find letters.").

Noticeably, all Grade 3 ELLs preferred keyboarding over handwriting. In most cases, the responses were related to the efficiency of keyboarding: "I can write more faster," "it's easier to write," "I like typing - so my hand doesn't get tired," and "because I can read letters easily when it's typed up."

Educators' Perceptions of ELLs' Writing Ability

As Table 16 summarizes, five ELL educators answered a series of semi-structured questions. In terms of the computer availability at school, all educators stated that not all ELL classes are equipped with computer devices. Accordingly, ELLs' computer use was limited to "online learning programs" they might encounter in their classes. The students also were rarely required to write or type stories with computers. Regarding the test mode preference, three out of five teachers (especially the Grades 1 and 2 teachers) preferred the handwriting tasks, while a Grade 3 teacher favored the Online-Keyboarding task.

Table 16. ELL Educator's Interview Responses

Educator	Computer Availability	ELL Computer Use	Preferred Writing Test Mode
Sara	Several computers in mainstream class	ELLs do not use computers in my class because we do not have the time	Paper because students (and teachers) can check what they wrote; easier to monitor during test administration
Terry	Computers in ELL class	ELLs use computers in my class and enjoy working on accelerated reading program	Online-Keyboarding task because children do well (they can type well) and these responses can be centrally scored
Amy	Several computers in mainstream class	ELLs (up to Grade 3) only need to log into the computer and know how to click. They work on "I-ready" – 15 minutes per session per child	Paper or Online-Handwriting because children can erase easily when mistakes are made
Delores	Several computers in mainstream class and four in ELL class	Grades 1–2 (just learning to use computers); G3 (need to type words). They use "imagine learning", "I-ready" –45 minutes per week per software	Depends on individual child's ability
Valerie	Several computers in mainstream class	Computer comfort depends on age of ELLs. They use "imagine learning"	Paper or Online-Handwriting for K-1; Online-Keyboarding for Grade 2 and up

Discussion and Conclusion

The main objective of the present study was to investigate the effects of test mode on young ELL children's writing performance. In response to calls for more refined analysis on children's actual writing (McDonald, 2002), this study explored not only the quantified test scores, but also the discrete features of composition quality produced by Grades 1–3 ELLs.

Overall, the delivery-response modes adopted in the present study had somewhat differential effects with regard to grade levels. Grades 1–2 ELLs received the highest scores on their writings in the Paper and/or Online-Handwriting tasks. Meanwhile, Grade 3 ELLs performed equally well regardless of the three administration modes. In addition, their differences in writing quality measures relative to the test modes appeared to be marginal compared to those of Grades 1–2 ELLs. This finding should be interpreted with some caution given the limited number of Grade 3 ELLs; yet little variability in *both* quantified and qualitative outcomes suggest that of all three grade levels, it was Grade 3 ELLs who were less likely to be affected by the test mode.

On the other hand, for Grades 1–2 ELLs, there was a significant effect of test mode on their overall test performance. More specifically, Grades 1–2 ELLs displayed noticeably better performance when being tested on paper, or at least in an administrative mode in which handwriting is allowed. Yet in terms of the test scores, a nuanced difference was observed between the two grade levels. Grade 1 ELLs' responses to the Paper task outscored each of the

texts produced via the two online-administrated modes. For Grade 2 ELLs, score differences were significant between the Paper and the Online-Keyboarding tasks; no measurable differences existed between the Paper and the Online-Handwriting tasks. However, this could indicate that while paper-based testing was the optimal condition for both grades, such a tendency was more transparent in Grade 1 ELLs' performance. It can be speculated that, they may have been better accustomed to writing on paper in general.

Such a strong tendency of the test mode effect on Grades 1 and 2 ELLs are somewhat less likely to be in line with previous comparability studies on writing test mode (e.g., Burke & Cizek, 2006; Wolfe et al., 1996; Wilmes et al., 2016). For instance, Grades 4 and 5 ELLs in Renn et al. (2015) performed comparably well in both handwriting and keyboarding test modes. Both Burke and Cizek (2006) and Wolfe et al. (1996) found a weak, conditional effect of test mode on older native-English speaking students. In Burke and Cizek (2006), the students with lower self-perceived computer skills performed significantly poorer on the computer-administered test. In a similar vein, Wolfe et al. (1996) found that the mode effect was only apparent with the students who had lower levels of computer familiarity. Without explicit measures of computer-related abilities tested on larger sample sizes, the present study cannot answer whether the results from Grades 1 and 2 are due to individual differences in computer use and test modes. Yet from the finding that the effect was particularly observed in younger children, it can be concluded that paper test mode may have been developmentally appropriate specifically for the children with developing composition and/or computer literacy.

Analyses of writing quality yielded somewhat similar patterns consistent with the quantitative results. The paper condition was favorable to the *fluency* measure, which was a strong tendency for Grades 1 and 2 ELLs. In particular, Grade 2 ELLs produced a greater volume of language as well as syntactic units in the Paper task. Such results corroborate findings from previous comparability studies. Berninger et al. (2009) revealed that elementary school children produced more words in handwriting tasks than in keyboarding tasks. Wolfe et al. (1996) additionally found that middle school students wrote more in the handwriting mode, subsequently being awarded higher scores for their handwritten texts.

In terms of *accuracy,* while the overall error rate was low (as evidenced from the average number of errors per 10 words), the average number of grammatical errors was highest in the Paper task and lowest in the Online-Keyboarding task for all grade levels. Use of *syntactic complexity*, especially regarding *cohesive devices* (e.g., reference markers) and *lexical diversity* (e.g., number of unique words) measures were higher on the Paper task. This pattern was more nuanced with Grade 2 ELLs, who consistently displayed greater use of the cohesive devices for the Paper task than the two online tasks. Somewhat in accordance with the *fluency* results, it could be speculated that lengthier texts (which contain more words) may display more complexity but also more inaccuracy of language usage. While no reference was made to the test modes, Bae (2001) found a strong connection between text length and the number of cohesive devices (particularly reference markers) with young ELLs used in the text ($r = 0.927$);

grammatical accuracy showed a moderate relationship with text length ($r = 0.650$). This finding led Bae to conclude a positive relationship existed between text length and writing quality.

Taken altogether, the present study's results seem to indicate that while fluent writers are not necessarily bound to accurate language usage (see Skehan [1998] for a discussion on the trade-off hypothesis among performance measures), their responses are likely to contain more occurrences of a variety of linguistic features. Notably, this study finds such a pattern among the ELLs responding in the Paper condition. This finding suggests that of the three modes, it is in the Paper condition that ELLs produced quality texts, which in turn, received higher ratings. However, the degree to which such qualitative differences accounted for the significant differences in test scores is not clear.

At the same time, the Online-Keyboarding condition may not have been ideal for younger children in the present study. One possible interpretation can be drawn from the observation of children's test-taking, which indicated that transcribing skills (e.g., handwriting, keyboarding) may have come into play. Previous researchers on early writers' handwriting vs. keyboarding performance drew on the cognitive-psychological models of writing (Hayes, 2006), which essentially put forth a positive association between transcribing skills (e.g., handwriting and keyboarding) and writing outcomes (Wollscheid, Sjaastad, & Tømte, 2016). According to this view, expertise in the lower-level mechanical skills in writing (e.g., keyboarding, handwriting) can free one's cognitive capacity for engaging in more advanced writing processes and skills (McCutchen, 1996); put differently, higher-order processes (e.g., planning, reviewing) associated with content generation may be bound to the mechanical demands of writing. If this were the case for the Grades 1 and 2 ELL children in the present study, their superior performance in the Paper task may indicate that they had more familiarity with handwriting than keyboarding to the extent of generating more structured, quality writing.

Yet regardless of the visible difficulties in typing, the majority of the same ELLs (and all Grade 3 ELLs) preferred keyboarding to handwriting tasks. Interestingly, these ELLs pointed out the efficiency in writing when keyboarding (e.g., speedier writing), which opposes the *fluency* results discussed above. For these ELLs, striking each key on the keyboard may have made their composing process much easier than handwriting letters one by one. Indeed, both Lee (2004) and Whithaus et al. (2008) demonstrated that college students, regardless of how they performed, tended to prefer the computer-delivered mode owing to the cumbersome nature of penmanship in editing and correcting when handwriting. In addition, for ELLs in this study, the novel aspect of keyboarding may have prompted them to be more eager to type their responses; this preference makes sense, considering their limited access to computers in the classrooms (see Table 15). As such, children's lack of typing skills did not limit how they *perceived* and *engaged* in the Online-Keyboarding task. In fact, children's responses imply that keyboarding could impose lesser mechanical demand than handwriting (Connelly et al., 2007). If sufficient and appropriate instruction on keyboarding is provided, ELLs' test-taking as well as writing experience in online testing could be made less challenging. According to the cognitive-psychological perspective of writing, it is plausible that engaging in constant practice on a specific transcription skill would

likely lead to enhanced writing performance on the corresponding response format (Christensen, 2004). In the classroom contexts, children should be sufficiently alerted to the ways the different administration modes are practiced and the type of skills that each mode requires (Burke & Cizek, 2006). Alongside such structured instructions on writing on computers, keyboarding could become a promising writing test mode even for this group of young children.

One major limitation of this study lies in the imbalanced sample size of all three grades; in particular, making any generalizations from the results found from the 10 third-graders. Moreover, the children were sampled from three school sites in which computer instruction was restrictively offered, which may further impede generalizing results to districts where ELLs have ample opportunities to use computers in their English as a second language classrooms. Additionally, the current study's findings imply the effects of additional moderating factors on writing performance such as computer familiarity, keyboarding skills, and ELLs' grade levels. Future studies need to take into account the direct relationship between the computer-related facets and individual differences to further advance current understanding of the mode effects on ELL children's test performance. Studies could also adopt more fine-grained analysis to measure the writing quality; for instance, children's writing fluency could be further captured with the use of Inputlog (Leijten & Van Waes, 2013), which provides outputs associated with writing process such as characters per minute, words per minute and so on.

Despite these limitations, the present study has its unique contribution to the field of young-test-takers or language assessment in general from its investigations into the test mode effect on young ELL children. Most importantly, the results raise awareness of the appropriateness of the testing practices targeting young children, especially in the computerized testing environment. The major finding of the present study is that the young ELL test-takers demonstrated both quantitative and qualitative differences in terms of the written language they produce based on both administration and response modes. In general, students are able to best show what they can do when they are able to handwrite their responses via a paper test mode rather than online modes.

References

Arnold, V., Legas, J., Obler, S., Pacheco, M. A., Russell, C., & Umbdenstock, L. (1990). Do students get higher scores on their word-processed papers? A study of bias in scoring hand-written versus word-processed papers. Unpublished manuscript, Rio Hondo College, Whittier, CA.

Bae, J. (2001). Cohesion and coherence in children's written English: Immersion and English-only classes. *Issues in Applied Linguistics, 12*(1), 51–88.

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31*(2), 241–259.

Berninger, V. W., Abbott, R. D., Augsburger, A., & Garcia, N. (2009). Comparison of pen and keyboard transcription modes in children with and without learning disabilities. *Learning Disability Quarterly, 32*(3), 123–141.

Bisschop, E., Morlaes, C., Gil, V, G., & Jiménez-Suárez, E. (2016). Fluency and accuracy in alphabet writing by keyboarding: A cross sectional study in Spanish-speaking children with and without learning disabilities. *Journal of Learning Disabilities, 50*(5), 1–9.

Blackhurst, A. (2005). Listening, reading and writing on computer-based and paper-based versions of IELTS. *Research Notes, 21*, 14–17.

Breland, H., Lee, Y., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts: Response mode analyses* (TOEFL Research Report No. RR-75). Princeton, NJ: ETS.

Burke, J. N., & Cizek, G. J. (2006). Effects of composition mode and self-perceived computer skills on essay scores of sixth graders. *Assessing Writing, 11*(3), 148–166.

Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing, 16*, 2011.

Choi, S. W., & Tinkler, T. (2002, April). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting.* In annual meeting of the National Council on Measurement in Education, New Orleans. Retrieved from https://www.researchgate.net/publication/274713232_Evaluating_comparability_of_paper-and-pencil_and_computer-based_assessment_in_a_K-12_setting_1

Christensen, C. A. (2004). Relationship between orthographic-motor integration and computer use for the production of creative and well-structured written text. *British Journal of Educational Psychology, 74*(4), 551–564.

Connelly, V., Gee, D., & Walsh, E. (2007). A comparison of keyboarded and handwritten compositions and the relationship with transcription speed. *British Journal of Educational Psychology, 77*, 479–492.

Crook, C., & Bennett, L. (2007). Does using a computer disturb the organization of children's writing? *British Journal of Developmental Psychology, 25*(2), 313–321.

Donker, A., & Reitsma, P. (2007). Young children's ability to use a computer mouse. *Computers and Education, 48*(4), 602-617.

Fashola, O. S., Drum, P. A., Mayer, R. E., & Kang, S.-J. (1996). A cognitive theory of orthographic transitioning: Predictable errors in how Spanish-speaking children spell English words. *American Educational Research Journal, 33*, 825–843.

Freeman, A. R., MacKinnon, J. R., & Miller, L. T. (2005). Keyboarding for students with handwriting problems: A literature review. *Physical & Occupational Therapy in Pediatrics, 25*(1/2), 119–147.

Green, T., & Maycock, L. (2004). Computer-based IELTS and paper-based versions of IELTS. *Research Notes, 18*, 3–6.

Hayes, J. R. (2006). New directions in writing theory. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 28–40). New York: Guilford Press.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 5*(2), 1–49.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (Research Report No. 3). Urbana, IL: National Council of Teachers of English.

Kröhne, U., & Martens, T. (2011). 11 Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft*, *14*(2), 169–186.

Laborda, J. G., Royo, T. M., & Bakieva, M. (2016). Looking towards the future of language assessment: Usability of tablet PCs in language testing. *Journal of Universal Computer Science, 22*(1), 114–123.

Lee, H. K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing, 9*(1), 4–26.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, *30*(3), 358-392.

Lindgren, E., & Stevenson, M. (2013). Interactional resources in the letters of young writers in Swedish and English. *Journal of Second Language Writing, 22*(4), 390–405.

Ling, G. (2017). Are TOEFL iBT writing test scores related to keyboard type? A survey of keyboard-related practices at testing centers. *Assessing Writing, 31*, 1–12.

McCarthy, P.M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2). 381-392.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8*, 299–325.

McDonald, A. S. (2002).The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. *Computers and Education, 39*, 299–312. http://dx.doi.org/10.1016/S0360-1315(02)00032-5

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in second language acquisition, 20*(1), 83–108.

Plakans, L., Gebril, A., & Bilki, Z. (2016). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing*, https://doi.org/10.1177/0265532216669537

Powers, D. E., Fowles, M. E., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*, 220–233.

Richardson, K., Calnan, M., Essen, J., & Lambert, L. (1976). The linguistic maturity of 11-year-olds: Some analysis of the written compositions of children in the National Child Development Study. *Journal of Child Language, 3*(1), 99–115.

Renn, J., DeMarco, N., & MacGregor, D. (2015, October). *Composing at the keyboard: An investigation into the effects of mode of response to writing tasks*. Paper presented at the annual meeting of the East Coast Organization of Language Testers, Washington, DC.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5*(3). Retrieved from: http://epaa.asu.edu/ojs/article/view/604/726

Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *Teachers College Record*. Retrieved from http://www.tcrecord.org/Content.asp?ContentID=10709

Russell, M., & Tao, W. (2004). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Practical Assessment, Research, & Evaluation, 9*(1), 1–9.

Shapiro, L. R., & Hudson, J. A. (1991). Tell me a make-believe story: Coherence and cohesion in young children's picture-elicited narratives. *Developmental Psychology, 27*(6), 960–974.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Taylor, C., Jamieson, J., Eignore, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (TOEFL Research Rep. No. 61). Princeton, NJ: ETS.

Weir, C. J., O'Sullivan, B., & Jin, Y. (2007). Does the computer make a difference? The reaction of candidates to a computer-based versus a traditional handwritten form of the IELTS writing component: Effects and impact. *IELTS Research Report, 7*(6). British Council and IELTS Australia Pty Limited. Retrieved from International English Language Testing System website: https://www.ielts.org/-/media/research-reports/ielts_rr_volume07_report6.ashx

Whithaus, C., Harrison, S. B., & Midyette, J. (2008). Keyboarding compared with handwriting on a high-stakes writing assessment: Student choice of composing medium, raters' perceptions, and text quality. *Assessing Writing, 13*, 4–25.

WIDA Consortium. (2012). *Development and Field Test of MODEL Grades 1-2 and 3-5*. Madison, WI: The Board of Regents of the University of Wisconsin System.

Wilmes, C., Olsen, P., & Montee, M. (2016, June). *Issues in keyboarding on assessments for English learners.* Paper presented at National Conference on Student Assessment, Philadelphia, PA.

Wolfe, E. W., Bolton, S., Feltovich, B., & Niday, D. M. (1996). The influence of student experience with word processors on the quality of essays written for a direct writing assessment. *Assessing Writing, 3*(2), 123–147.

Wolfe, E. W., & Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning & Technology, 8*(1), 53–65.

Wollscheid, S., Sjaastad, J., & Tømte, C. (2016). The impact of digital devices vs. pen(cil) and paper on primary school students' writing skills: A research review. *Computers & Education, 95*, 19–35.

Appendix A. Observation Notes

Instruction: Select 1-2 students from the group per session. Take notes of any noticeable behavior while students complete each writing task. For example, the student may be struggling to keyboard and therefore cannot complete the task.

Date: _____ Location: _____ School: _____

Student Name or ID: _____ Grade: _____

Task sequence (indicate number 1, 2, or 3)			Observation notes		
Paper	Online-Hand write	Online-Keyboard	Paper task	Online-Handwriting task	Online-Keyboarding task

Appendix B. Post-test Questions for Grades 1–3 ELLs

Post-test questions for Grades 1–3 ELL children:

- What was your favorite task/activity? Why did you like that one?
- Do you prefer handwriting your task/activity or typing? Why?

Appendix C. Semi-structured Interview Questions for Grades 1–3 ELL Teachers

Semi-structured Interview Questions for Grades 1–3 ELL Teachers

Instruction: Ask teachers the following questions (questions will include, but will not be limited, to the below).

Background Information

- Name:
- State:
- District:
- School:
- Position:
- Among Grades 1–3 ELLs, what grade levels do you support? (select all)
    - Grade 1
    - Grade 2
    - Grade 3
- How many ELLs do you support?
- How many years of experience do you have working with ELLs?

Currently, there are three options for completing a writing task:

- Paper: students complete their writing test using a paper booklet
- Online task with handwritten response: students read the tasks online and hand write their response on paper
- Online task with keyboarded response: students read the tasks online and keyboard their response

Questions for Educators

- What is the most appropriate method for assessing ELLs' writing ability? Also, please explain why.
    - paper
    - online task with handwritten response
    - online task with keyboarded response
    - please explain why:
- How comfortable can ELLs keyboard/type in general? (some examples below)
    - They cannot type
    - They can find and type most of the letters in the alphabet
    - They can type simple words
    - They can type in full sentences
    - Other (please specify): _____
- How often are ELLs required to work on a computer during instructional hours? (some examples below)
    - 1-2 times/week
    - 3-4 times/week
    - everyday
    - Other (please specify): _____
- For what purpose are Grades 1–3 ELLs required to work on a computer during instructional hours?
- What keyboarding support/instruction is offered to Grades 1–3 ELLs at school?

Appendix D. Writing Scoring Rubric

Level	Linguistic Complexity	Vocabulary Usage	Language Control
6 Reaching*	A variety of sentence lengths of varying linguistic complexity in a single tightly organized paragraph or in well-organized extended text; tight cohesion and organization	Consistent use of just the right word in just the right place; precise Vocabulary Usage in general, specific or technical language.	Has reached comparability to that of English proficient peers functioning at the "proficient" level in state-wide assessments.
5 Bridging	A variety of sentence lengths of varying linguistic complexity in a single organized paragraph or in extended text; cohesion and organization	Usage of technical language related to the content area; evident facility with needed vocabulary.	Approaching comparability to that of English proficient peers; errors don't impede comprehensibility.
4 Expanding	A variety of sentence lengths of varying linguistic complexity; emerging cohesion used to provide detail and clarity.	Usage of specific and some technical language related to the content area; lack of needed vocabulary may be occasionally evident.	Generally comprehensible at all times, errors don't impede the overall meaning; such errors may reflect first language interference.
3 Developing	Simple and expanded sentences that show emerging complexity used to provide detail.	Usage of general and some specific language related to the content area; lack of needed vocabulary may be evident.	Generally comprehensible when writing in sentences; comprehensibility may from time to time be impeded by errors when attempting to produce more complex text.
2 Beginning	Phrases and short sentences; varying amount of text may be copied or adapted; some attempt at organization may be evidenced.	Usage of general language related to the content area; lack of vocabulary may be evident.	Generally comprehensible when text is adapted from model or source text, or when original text is limited to simple text; comprehensibility may be often impeded by errors.
1 Entering	Single words, set phrases or chunks of simple language; varying amounts of text may be copied or adapted; adapted text contains original language.	Usage of highest frequency vocabulary from school setting and content areas.	Generally comprehensible when text is copied or adapted from model or source text; comprehensibility may be significantly impeded in original text.

Appendix E. Coding Scheme for Qualitative Analysis

Category	Sub-Category	Description	Unit/Calculation
Task completion		Number of pictures described in each response	Frequency count of the number of sample words used from the Word Bank in each response (Max. 8 sample words)
		Number of words used from the Word Bank in each response	Frequency count of the number of pictures mentioned in each response (Max. 5 pictures)
Fluency		Number of total words produced in each response	Total number of words in each response
Accuracy			
Lexical error	Homophones	Words that have the same sound as another word but are spelled differently and have a different meaning (e.g., *their*/*there*, *red*/*read*, *too*/*to*, *than*/*then*)	Frequency count of the instances
	Words with similar spelling, but with different meaning	Words that have similar spelling to the intended word but convey the wrong meaning (e.g., *leader*/*ladder*, *dawn*/*down*, *want*/*went*, *sale*/*sell*)	
	Awkward word choices	Words that seem to be a misfit to the context (e.g., He *put* the lemonade in glass)	
	Missing words	Words that are missing (e.g., Gave back to ___)	
Grammatical error	Part-of-speech errors Morphological errors	Incorrect usages of articles, verbs, pronouns Morphological errors (past tense –ed, possessive 's, 3rd person singular –s)	
Errors per 10 words		Number of errors (with respective to each error category) per 10 words	[Total number of errors / Total number of words] * 10
Complexity			
Lexical variety	Unique words used (type/token frequency)	The unique number of words used vs. the total number of words used in each response	
	Word frequency	Five most frequently used words in each response	Instances of the top five frequent words and frequency counts of each word
Syntactic complexity	Number of T units	Total number of T units used in each response	Total number of T units
	Mean T unit	Average length of T units used in each response	Total number of T units / Total number of words
	Mean sentence length	Average length of each sentence in each response	Number of words per sentence
	Cohesion	Reference markers: Pronominal (*he, her, they, theirs*), proper noun (*Max, Susan*), demonstrative (*this/these*, *that/those*, *there*) Conjunctive markers: Additive (*and*, *or*), adversative (*but*), temporal (*then*, *and then*, *next*, *finally*)	Frequency count of reference and conjunctive markers