# Impact of Computer-Based Administration of Writing Tests on Elementary Grade Students' Writing

Dale J. Cohen[1], Jon Cohen[2], Irene Hunting[3], and Heather O'Neill[2]

[1]University of North Carolina Wilmington

[2]American Institutes for Research

[3]Arizona Department of Education

## Author's Note

*Authors are listed in alphabetical order.*

Dale J. Cohen, Department of Psychology, University of North Carolina Wilmington, 601 South College Road, Wilmington, NC 28403-5612; Jon Cohen and Heather O'Neill, American Institutes for Research, 1000 Thomas Jefferson Street, NW, Washington, DC 20007. Irene Hunting, Arizona Department of Education, State Government Office, 1535 W Jefferson St, Phoenix, AZ 85007. Please address correspondence to Dale J. Cohen, Ph.D. Email: cohend@uncw.edu. Phone: 910.962.3917. Fax: 910.962.7010. This publication is intended to promote the exchange of ideas among researchers and policy makers. The views expressed in it

are those of the authors and do not necessarily reflect the position of the American Institutes for Research.

## Abstract

In the field, educators tend to fear that keyboarding demands will lower the quality of student writing on large scale state assessments. Specifically, the added burden of keyboarding overcome any positive influence of computer use and this will prevent students from demonstrating their writing skill. The extant research on this issue has shown mixed results. Here, we present a study that directly compares students' hand-written drafts to their computer input final versions on an operational state test. We found that students tended to improve their drafts in terms of structure and content when transferring the paper rough draft to the computer. Overall, these findings support the use of online writing assessment.

# Impact of Computer-Based Administration of Writing Tests on Elementary Grade Students' Writing

With the growing availability of the technology, computers are increasingly being used for high-stakes writing assessments. The switch to computer-based writing assessments from paper-based writing assessments raises the question of whether the mode of administration (hand-written on paper vs computer input) influences students' writing performance. The data in the extant literature is mixed, with some research showing that computers aid writing performance (e.g., Goldberg, Russell, & Cook, 2003; Russell, & Plati, 2001), some showing no influence of computers (e.g., Mogey & Hartley, 2013; White, Kim, Chen, & Liu, 2015), and others showing that computers interfere with writing, at least for some subgroups (e.g., Chen, White, McCloskey, Soroui, & Chun, 2011). In the field, many educators fear that keyboarding demands will lower the quality of student writing, and that the added burden will prevent students from demonstrating their writing skill.

In general, the influence of mode of administration on adults' writing performance appears to be qualitatively different from that of children. Specifically, writing on the computer appears to have more variable, and sometimes negative, effects on adults. For example, Chen, et al. (2011) directly assessed how the mode of test administration influenced the writing performance of adults. The authors asked 1607 adults (aged 16 and older), representative of the US population, to complete one of three writing tasks in a field test for the National Assessment of Adult Literacy (NAAL). Approximately half the volunteers completed the writing assessment on the computer. The authors found that, overall, writing performance was better when the test mode was on paper. However, there was an interaction between race-ethnicity (Black vs White), age (over 65 vs under 40), and employment status (unemployed vs employed) and test mode.

Here, some groups did more poorly than the population as a whole when writing on the computer. The authors suggested that these interactions may be explained by the overall finding that writing performance on the computer was positively correlated with computer experience.

Unlike adults, writing on a computer appears to have either no influence or a positive influence on children's writing performance. For example, White et al. (2015) conducted a large scale study comparing over 10,000 fourth grade students' computerized 2012 NAEP writing assessment responses to the 2010 NAEP fourth grade paper-based writing assessment. Overall, the results showed that students performed better on the computerized writing assessment than the paper-based assessment. However, this result is qualified by the finding that while high performing students did better on the computerized assessment, there was no difference in the writing performance of low and middle performing students. The authors also found a relation between prior exposure to writing on the computer and writing performance. These findings echo those of Horkay, Bennett, Allen, Kaplan, and Yan (2006) who analyzed over 4000 eighth grade students' performance on a paper and computerized version of the 2002 NAEP writing assessment. Overall, the authors found little or no influence of mode of test administration.

The findings cited above conflict a bit with earlier research that demonstrates an almost universal positive effect on children's writing performance. For example, Goldberg et al. (2003) conducted a meta-analysis on 23 studies that compared computer vs paper-based writing performance of K-12 students. The analysis showed that students' writing quality and quantity was improved on the computer relative to the hand-written responses. The authors conclude the computerized writing tools aid students' writing performance.

The influence of computers on writing performance of children appears to be either positive or negligible. Nonetheless, state testing programs often hesitate to ask younger students

to take writing tests online, with some states introducing online writing tests only in grade 5 and later. This paper explores whether this hesitation is warranted.

Here, we present a study that directly compares students' hand-written drafts to their computer input final versions on an operational state test. In most online writing assessments, students are provided with scratch paper on which to draft their responses, so studies of online writing often reflect this mixed-mode generation of text (Smarter Balanced Assessment Consortium, 2016; Pearson Education, Inc., 2016; Arizona Department of Education, 2016). The existence of early paper drafts allows us to examine changes in both content and writing conventions between the paper draft and the online final version. This allows us to examine whether students lose or improve content when creating the online draft, and the extent to which they edit their work. The main question we ask is whether elementary students are able to improve upon early, paper drafts when taking an online writing test.

## Current Study

This study draws on data collected from a single elementary school in Arizona. This was the first statewide assessment administered online in this school. The school's writing curriculum emphasized the writing process, including careful planning and early drafts. As a result, most of the students produced substantial drafts on paper before entering or composing their final drafts on the computer. Matching the paper pre-work to the final computer-based essays made it possible to trace the progression of ideas and writing from the early drafts to the final draft, which allowed identification and analysis of changes, additions, omissions, and other characteristics of the paired drafts and final versions. These essays comprised the basis for the current study, which addresses the following research questions:

- Does computer-based response enhance or impede young students' compilation of relevant information into an essay?
- Does computer-based response enhance or impede young students' grammar and spelling?
- Does computer-based response enhance or impede young students' structuring of their essay?

Elementary students in Arizona are expected to write a variety of types of texts for a variety of purposes. Their writing should reflect information gathered from multiple print and other sources and integrate this information while avoiding plagiarism. Students are expected to demonstrate an understanding of the subject under investigation, and produce clear and coherent work. The organization and structure of their essays should be tailored to the audience and purpose (Arizona Department of Education, 2013).

In keeping with these expectations, the writing assessment begins with several readings on a common topic. Students are given a question to address and are scored on their ability to structure their argument, draw information from the sources provided in a coherent way to support their points, and abide by accepted writing conventions. Students have access to dictionaries and thesauri throughout the writing assessment. The online assessment includes some typical word processing tools like copy, cut, and paste, as well as dictionaries and thesauri. Students do not have access to spell check and grammar check tools.

The participating school is located in the Phoenix area and includes grades K through 5. It is a Title I school which means it receives financial assistance for having a high percentage of children from low-income families. The school's mission statement mentions that the school has

computers in all classrooms as well as two Mac Labs. The in-class writing instruction employed by the school is titled 'Write from the Beginning and Beyond'. This writing process teaches students to use thinking maps to organize their thoughts. Then, from the maps or graphic organizers, students write a rough draft of their essay. After revision, the students produce a final version. Most of the pre-work gathered for this study includes both a graphic organizer and a rough draft in paragraph form.

## Method

### Participants

Three-hundred and fifteen students from the studied school participated in writing assessments in grades 3-5. Of these, 250 were suitable for inclusion in the study. Table 1 summarizes the disposition of each case. The largest group of students excluded from the study were 42 for whom preliminary work was not available, either because it was not collected or the students in question did no pre-work on paper. The next largest category of excluded students included 17 who provided little or no original text, including some students who simply copied parts of the reading passages that served as stimuli for the writing prompts. A few additional students provided responses unrelated to the prompt or responded in a language other than English.

The sample of 250 students was approximately evenly split among the grades 3, 4, and 5, with 84, 76, and 90 students respectively. A summary of the student scores on their final essays appears in Table 2.

**Procedure**

Consistent with the traits that the essays were designed to measure, the pre-work and final essays were coded to identify three types of information:

- *propositions*, which represent the content and evidence brought to bear in the service of the papers' theses;
- *grammar and spelling errors*, which correspond to the writing conventions on which the students were scored; and
- *essay structure*, which reflects the progression and presentation of content in the essay.

We discuss the coding of the student work for each of these traits below.

**Propositions.** We define a proposition to be a single thought or idea presented in a text. For example, the statement "The dog chased a cat and a squirrel," represents two distinct propositions: a) "the dog chased a cat," and b) "the dog chased a squirrel."

Coders with a minimum of a four-year degree were recruited and trained to identify propositions in student writing. Coders worked with a *draft/final pair*, a paper first draft or other pre-work on paper and the final essay the student composed or entered into the computer. The coders numbered and inventoried specific propositions in each document in a draft/final pair. Each draft/final pair was independently coded by two coders, with one coding the draft first and the final second, and the other coding in reverse order.

Each inventoried proposition was categorized as "on target," or "off target." Consistent with the scoring rubrics for the essays, on-target propositions were those that were relevant to the topic and a) either stated or implied in the stimulus material, or b) reflected background knowledge that was relevant and supporting the main theme of the essay.

Results from the two coders were compiled and discrepancies reconciled by a supervisor. This process resulted in four pieces of information per case:

- Number of on-target propositions found in the pre-work;
- Number of off-target propositions found in the pre-work;
- Number of on-target propositions found in the final essay; and
- Number of off-target propositions found in the final essay.

**Grammar and spelling.** The same group of coders coded each essay for grammar and spelling errors. Grammar errors were limited to those that were grade appropriate. For example, misusing commas and periods, misspelling words, or using an incorrect homonym were counted as errors. Misuse of semicolons, colons, and advance phrase structures were not. This was consistent with the Conventions scoring rubric, which reflected the Arizona state writing standards.

As with the propositions, spelling and grammar were each coded independently by two raters, and discrepancies resolved by one of the supervisors.

**Structure.** Coders were trained in the scoring rubric designed to evaluate the "Structure and Organization" of the essays, and were asked to make a holistic judgment about the draft/final pairs. As with the other codings, each pair was coded twice independently, and discrepancies resolved by one of the supervisors.

## Data

The coding of propositions resulted in the data summarized in Table 3. Nominal increases in both on-target and off-target propositions are apparent in the final essays, indicating that, on average, students added to their essays as they composed their final draft.

The coding was acceptably reliable. We achieved excellent inter-rater reliabilities for the total number of propositions and the number of on-target propositions. Coders had more trouble classifying off-target proposition, with many more discrepancies resolved by our supervisors. Several of our coders tended to classify propositions as on-target if they were in any way on the correct topic. The rubrics, however, required that the propositions be related to the readings. This type misclassification was the most commonly corrected by our supervisors. Table 4 displays these data.

Table 5 summarizes the counts of grammatical errors, which also increased between the drafts and the final versions. The Pearson's inter-rater correlation here was .56 and .57 respectively for grammar errors in pre-work and errors in the final draft. Correlations with the final resolved score were .79 and .82.

Table 6 summarizes the changes in structure from the pre-work to the final work. There were 56 cases where the structure comparison between the pre-work and the final response was characterized as 'indeterminate'. For these responses, there was no written rough draft. The pre-work was limited to only a graphic organizer. Most of the graphic organizers consisted of a written introduction, followed by 3 main ideas with bulleted supportive information from the passage, and a written conclusion. There were a few students that drew a web with their main idea in the center with supporting information from the passage stemming from the main idea. Since these are graphical representations of the students' ideas, the effect on the structure of the response from writing on the computer could not be determined for these cases. For those that were coded the raters matched in 86.4% of cases.

While not perfect, the measures each had acceptable reliability. A series of ordered probit analyses (presented in Appendix A as Tables A1-A3) supported the validity of the final

measures. We predicted the final score on each dimension based on the final coded scores of the final essay. In each case, the number of on-target propositions was positively and significantly associated with the score on the dimension. These associations were strongest for the Evidence and Elaboration and Structure and Purpose, and weakest for the Conventions dimension, as should be expected.

The number of off-target propositions was negatively and significantly associated with the Evidence and Elaboration score and the Structure and Purpose score. The number of spelling and grammar was negatively associated with scores on all dimensions, with the association strongest for the Conventions dimension.

A pair of dummy variables indicated whether the Structure and Organization classification improved (variable 1) or degraded (variable 2), with all other cases coded as zero. These dummy indicators were not a significant predictor of any of the dimension scores; however, a coded decline in structure was associated with a nominal decrease in score and coded increase in structure was associated with a nominal increase. Given the small number of changes, and the second-order nature of the measure itself, we take this as evidence supporting the validity of our coders' judgments.

## Results

Table 7 presents the correlations among the proposition and grammar metrics. Focusing first on the same measures from pre-work to final essay, we see quite substantial correlations. The correlations for the proposition metrics between the pre-work and final are .88 (on-target propositions) and .90 (off-target propositions). These suggest that the responses changed, but not dramatically in terms of content covered. The correlation in the number of spelling and grammar errors between the pre-work and final is somewhat lower at .68, but still quite high.

Table 7 also reveals a negative correlation between the number of on-target and off-target propositions, suggesting that students who include more on-target propositions are less likely to also offer irrelevant information.

Do students responding on the computer lose information, or are they able to enhance their responses? Table 8 speaks directly to this question. Table 7 presents the *net change in on-target propositions*. This is calculated by subtracting the change in the number of off-target propositions from the change in the number of on-target propositions. For example, imagine a student who includes 5 on-target propositions and zero off-target propositions in his or her pre-work. In the hypothetical final essay we find 7 on-target propositions and one off-target proposition, for a net change of one (7-5 – (5-0) = 1). This serves as a summary measure of the improvement of the content of the essay.

Every grade showed a positive net change in on-target propositions. This was statistically significant at conventional levels in every grade but grade 4, and grade 4 trended in the right direction. Overall, students were able to improve the content of their essays when they composed or entered their final draft on the computer. As most elementary educators would expect, this was most pronounced at the fifth grade, where final work tend to diverge more from drafts.

A different story begins to emerge when we ask the same question about grammar and spelling. Table 9 presents the net change in the number of grammar and spelling errors from the pre-work to the final draft. Overall, students experience an increase in the number of grammar and spelling errors when composing or entering their final essays in the computer. Somewhat surprisingly, this effect is most pronounced at grade 5.

Table 10 sheds some light on this finding. Recall from the analysis of propositions that students tended to add information in their final drafts. When we look at the change not in a simple count of spelling and grammar errors, but as a ratio of such errors to the number of propositions, we are drawn to a somewhat different conclusion. Only in grade 5, where students added the most propositions between draft and final, does the change in the ratio approach statistical significance. We revisit this topic in our discussion.

A final draft also offers an opportunity to improve the structure of the response. Table 11 presents a statistical test of the data presented in Table 6. In grades 4 and 5 we find significant improvements in structure—significantly more students improved the structure of their essays than declined. The positive finding in grade 3 falls just shy of the .05 level. In the lower grades we note that changes in structure and organization were rare.

## Discussion

This study evaluated the impact of computer-based writing assessment on student essays. In particular, we looked specifically at whether the computer-response mode enhanced or degraded students' compilation of relevant material in an essay, their grammar and spelling, and the structure of their essays. Drawing on data collected from a single elementary school, we systematically compared paper-based rough drafts with the final drafts entered on the computer.

We found that students tended to improve their drafts in terms of structure and content when transferring the paper rough draft to the computer. Students were more likely to include additional relevant information, eliminate irrelevant information, and improve the structure and flow of their arguments. We saw no improvement in their grammar or spelling and saw some potential evidence of degradation. Review of the student drafts and final essays revealed a tendency to introduce typographical errors.

Overall, these findings support the use of online writing assessment. These findings are consistent with White et al. (2015), who also found that student essays were improved when entered online. This study focused on elementary students (grades 3-5), providing evidence that even these young children are currently able to respond online and improve their answers from draft form.

Our finding of a lack of improvement in grammar and spelling warrants further research. We hypothesize that students adding information during their online draft tend not to revisit the new material for editorial review. If this is the case, it is entirely reasonable to continue to base their scores on these conventions. However, if students are introducing new typographical errors simply due to developing keyboarding skills, assessment organizations might consider exempting such errors from the scoring rubrics.

While this study benefited from the use of a real-world assessment associated with accountability stakes, it is limited by its convenience sample. The data were drawn from a single, albeit diverse, elementary school. The generalizability of findings based on such samples is always tenuous. We believe that these findings will hold up in a large, representative, scientific sample—a very promising next step in this research.

## References

AzMERIT Test Administration Directions (2016). Phoenix, AZ: Arizona Department of Education. 14.

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*(1), 49–71.

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *The Journal of Technology, Learning and Assessment, 2*(1).

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), n2.

Mogey, N., & Hartley, J. (2013). To write or to type? The effects of handwriting and word-processing on the written style of examination essays. *Innovations in Education and Teaching International, 50*(1), 85–93.

PARCC Test Administration Manual for Computer-Based Testing (2016). Upper Saddle River, NJ: Pearson Education, Inc. 13.

Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a statemandated writing assessment. TC Record.*org. Available online at: http://www.tcrecord.org/Content. asp*.

Smarter Balanced Online Test Administration Manual (2016). Los Angeles, CA: Smarter Balanced Assessment Consortium. TAM-37.

White, S., Kim, Y. Y., Chen, J., & Liu, F. (2015). Performance of Fourth-Grade Students in the 2012 NAEP Computer-Based Writing Pilot Assessment: Scores, Text Length, and Use of Editing Tools. Working Paper Series. NCES 2015-119. *National Center for Education Statistics*.

Table 1

*Summary of cases included and excluded from study*

| Case Disposition | Number of Cases |
|---|---|
| **In Study** | **250** |
| Excluded, no pre-work | 42 |
| Excluded, response off topic | 1 |
| Excluded, insufficient text | 17 |
| Excluded, response not in English | 5 |
| **Total** | **315** |

Table 2

*Sample size and average writing scores of study sample by grade*

| **Grade** | **Number of Cases** | **Average Score, Structure and Purpose (1-4)** | **Average Score, Evidence and Elaboration (1-4)** | **Average Score, Conventions (0-2)** |
|---|---|---|---|---|
| All | 250 | 1.8 | 1.7 | 1.3 |
| Grade 3 | 84 | 2.2 | 2.1 | 1.3 |
| Grade 4 | 76 | 1.6 | 1.5 | 1.2 |
| Grade 5 | 90 | 1.7 | 1.6 | 1.3 |

Table 3

*Number of Propositions, on and off target, in pre-work and final versions, by grade*

| | On Target Pre-work Propositions | | On Target Final Propositions | | Off Target Pre-work Propositions | | Off Target Final Propositions | |
|---|---|---|---|---|---|---|---|---|
| | Average | SD | Average | SD | Average | SD | Average | SD |
| All | 9.8 | 6.64 | 11.1 | 7.16 | 2.4 | 3.81 | 2.7 | 3.74 |
| Grade 3 | 7.0 | 4.85 | 7.8 | 5.04 | 3.3 | 4.33 | 3.4 | 4.46 |
| Grade 4 | 7.6 | 5.48 | 8.3 | 5.62 | 2.8 | 3.92 | 3.2 | 3.99 |
| Grade 5 | 14.3 | 6.69 | 16.7 | 6.61 | 1.3 | 2.83 | 1.6 | 2.26 |

Table 4

*Interrater reliabilities in the coding and classification of propositions*

| Number of propositions | Correlation with second rater | Average correlation with final resolved score |
|---|---|---|
| Total number of propositions, Pre-work | .77 | .80 |
| Total number of propositions, Final | .75 | .78 |
| On target propositions, pre-work | .62 | .79 |
| On target propositions, final | .59 | .77 |
| Off target propositions pre-work | .21 | .53 |
| Off target propositions, final | .19 | .53 |

Table 5

*Number of grammar errors in pre-work and final versions, by grade*

| | Grammar errors, Pre-work | | Grammar errors, Final | |
|---|---|---|---|---|
| | Average | SD | Average | SD |
| All | 13.9 | 9.56 | 16.6 | 10.54 |
| Grade 3 | 13.2 | 8.49 | 14.5 | 8.81 |
| Grade 4 | 13.2 | 9.64 | 14.6 | 8.71 |
| Grade 5 | 15.1 | 10.38 | 20.4 | 12.34 |

Table 6

*Number of cases changing structure and organization from pre-work to final work, by grade*

| | Improved | No Change | Deteriorated | No Determination Possible |
|---|---|---|---|---|
| All | 31 | 159 | 4 | 56 |
| Grade 3 | 6 | 49 | 1 | 28 |
| Grade 4 | 10 | 60 | 2 | 4 |
| Grade 5 | 15 | 50 | 1 | 24 |

Table 7

*Correlations among measurements among student pre-work and final work*

| Metric | On Target Pre-work Propositions | On Target Final Propositions | Off Target Pre-work Propositions | Off Target Final Propositions | Spelling and Grammar Errors, Pre-work | Spelling and Grammar Errors, Final |
|---|---|---|---|---|---|---|
| On Target Pre-work Propositions | 1.00 | | | | | |
| On Target Final Propositions | 0.88 | 1.00 | | | | |
| Off Target Pre-work Propositions | -0.36 | -0.37 | 1.00 | | | |
| Off Target Final Propositions | -0.36 | -0.38 | 0.90 | 1.00 | | |
| Spelling and Grammar Errors, Pre-work | 0.24 | 0.16 | 0.11 | 0.09 | 1.00 | |
| Spelling and Grammar Errors, Final | 0.23 | 0.32 | 0.05 | 0.09 | 0.68 | 1.00 |

Table 8

*Net change in on-target propositions from pre-work to final work*

| | Mean | SD | t-statistics | p>\|t\| |
|---|---|---|---|---|
| All | 1.1 | 4.00 | 4.18 | 0.00** |
| Grade 3 | 0.6 | 2.56 | 2.09 | 0.04* |
| Grade 4 | 0.3 | 1.56 | 1.47 | 0.15 |
| Grade 5 | 2.2 | 5.88 | 3.48 | 0.00** |

* significant at $\alpha = .05$; ** significant at $\alpha.01$.

Table 9

*Net change in grammar errors from pre-work to final work*

| | Mean | SD | t | p>\|t\| |
|---|---|---|---|---|
| All | 2.7 | 8.14 | 5.28 | 0.00** |
| Grade 3 | 1.3 | 8.14 | 1.43 | 0.16 |
| Grade 4 | 1.4 | 5.95 | 2.00 | 0.02* |
| Grade 5 | 5.3 | 9.14 | 5.48 | 0.00** |

* significant at $\alpha = .05$; ** significant at $\alpha .01$.

Table 10

*Ratio of grammar errors to propositions in pre-work and final work*

| | Pre-work | Final Work | Difference | t (Difference) | p>\|t\| |
|---|---|---|---|---|---|
| All | 1.30 | 1.33 | 0.03 | 0.72 | 0.475 |
| Grade 3 | 1.46 | 1.48 | 0.02 | 0.20 | 0.841 |
| Grade 4 | 1.33 | 1.29 | -0.04 | -0.56 | 0.574 |
| Grade 5 | 1.13 | 1.24 | 0.11 | 1.94 | 0.056 |

* significant at $\alpha = .05$; ** significant at $\alpha.01$.

Table 11

*Structure and organization: number of essays improving less the number of essays deteriorating*

| | Improved – Deteriorated | t | p>\|t\| |
|---|---|---|---|
| All | 27 | 4.82 | 0.000** |
| Grade 3 | 5 | 1.93 | 0.058 |
| Grade 4 | 8 | 2.38 | 0.020* |
| Grade 5 | 14 | 3.85 | 0.000** |

* significant at $\alpha = .05$; ** significant at $\alpha .01$.