CALA Report 208

Effects of Handwritten Versus Computer-Written Modes of Communication on the Quality of Student Essays

FJ King
Faranak Rohani
Carol Sanfilippo
Natalee White

September 2008

Center for Advancement of Learning and Assessment
Florida State University, Tallahassee, FL
www.cala.fsu.edu

Center for Advancement of Learning and Assessment (CALA)—As a pioneer in research and the design of multimedia instructional materials and customized assessments, CALA assists policy makers and educators by providing them analyses and practical solutions.

Editors: Karen Hawkins, Alice Fisher

Senior Graphic Designer: Colin Dwyer

Media Specialist: Liane Schuessler

Software Engineers: Sean Coleman, Michael Davidson, Thomas Bonfield, Gagan Chhatwal

Center for Advancement of Learning and Assessment

Florida State University

210B Sliger Building

2035 E. Paul Dirac Drive

Tallahassee, FL 32306-2801

Phone: (850) 645-CALA (2252)

Fax: (850) 645-7581

www.cala.fsu.edu

Abstract

This study examined the effects of mode of writing on students' preplanning processes and essay scores. Graduate and undergraduate students in Florida's college and university education programs were given prompts released from the Florida Teacher Certification Examinations. Students wrote to one prompt using paper and pencil and to another using the computer. Students' preplanning activities were captured in both modes. In addition, a questionnaire was administered to determine students' writing processes and their familiarity with computers. Approximately half of the essays in each mode were transcribed. The original and transcribed essays were then scored holistically. Findings showed no significant difference in scores between the modes, nor was rater bias present. Students tended to do more preplanning for the paper-and-pencil essays than for the computer-based essays; however, there was little correlation between preplanning and the resulting scores for the computer-based essays.

Effects of Handwritten Versus Computer-Written Modes of Communication on the Quality of Student Essays

The state of Florida requires, by law, that persons seeking to become teachers pass the Florida Teacher Certification Examinations (FTCE) for Professional Education, General Knowledge, and the subject area(s) of specialization. Included in the General Knowledge examination are four subtests, one of which requires examinees to write an essay. From 2004–2007, more than 68,000 examinees took the essay portion of the FTCE for the first time. Given the national trend away from paper-and-pencil assessment toward computer-based assessment, the Florida Department of Education (FDOE) requested that the Center for Advancement of Learning and Assessment at Florida State University (FSU/CALA) conduct research related to the effects of the mode that examinees use to respond to an essay prompt. The purpose of the research is twofold:

1. to determine if subjects' scores indicate a difference in quality of essays when they respond to an essay prompt using paper and pencil versus computer.
2. to determine if there are systemic variations in the raters' scoring of the essays produced using paper and pencil versus computer.

This study includes a review of literature related to the cognitive and writing processes through which a writer moves to produce an essay; it also includes literature reviews related to the scoring of essays. The empirical study itself includes comparisons of essay scores written by FTCE candidates in both composition modes as well as a compilation of information from the examinees about several experiential and affective variables.

Review of Literature

Research related to the process of writing has been concerned primarily with how students can best be taught to write well and how to judge the quality of student writing. Understanding the various components of this complex task has occupied several scholars. The first section of this review provides literature describing the psychological constructs of writers and the writing process. The second section covers studies in which the quality of paper-and-pencil-produced (handwritten) essays was compared with the quality of computer-produced essays. Since researchers in this field have worked with groups of subjects who varied in age, status (students or nonstudents), experience with computer technology, and other dimensions, the focus here will be on studies that emphasize the relative quality of handwritten versus computer-produced essays in groups of high school or college-age students. In addition, reported results of these studies will be reviewed to determine whether increasing availability of computer technology to students has influenced the perceived quality of essays produced in the two different modes.

Psychological Constructs and the Writing Process

Cronbach and Meehl (1955), in their classic paper "Construct Validity in Psychological Tests," defined a psychological construct as "some postulated attribute of people," giving "cheerfulness" and "anxiety proneness" as examples. For a writing task, this set of psychological "baggage" determines the outcome of the product. Among the psychological constructs that are related to the quality of a writer's work are both general and specific abilities related to the composing process; less obvious are the writer's attitudes and personal characteristics, such as persistence or ambition. In the Hayes and Flower Writing Model, as modified by Ackerman and Smith (1988), psychological constructs (e.g., verbal ability, ability to organize, ability to set goals) are implied in the process labels used: environment (unconscious input), the writer's long-term

memory, and formal training (conscious input). Taking into account the reality of writing as a task, the model also includes the physical environment and assignment attributes as factors influencing the process (see Figure 1).

Three basic skills are included in the performance model: planning, translating, and reviewing. Planning comprises three subskills: generating ideas, organizing ideas, and setting goals. Translating these plans into a written document involves converting ideas into grammatically correct sentences and paragraphs. The final process of reviewing is carried out to improve the quality of the document and eliminate mechanical errors.

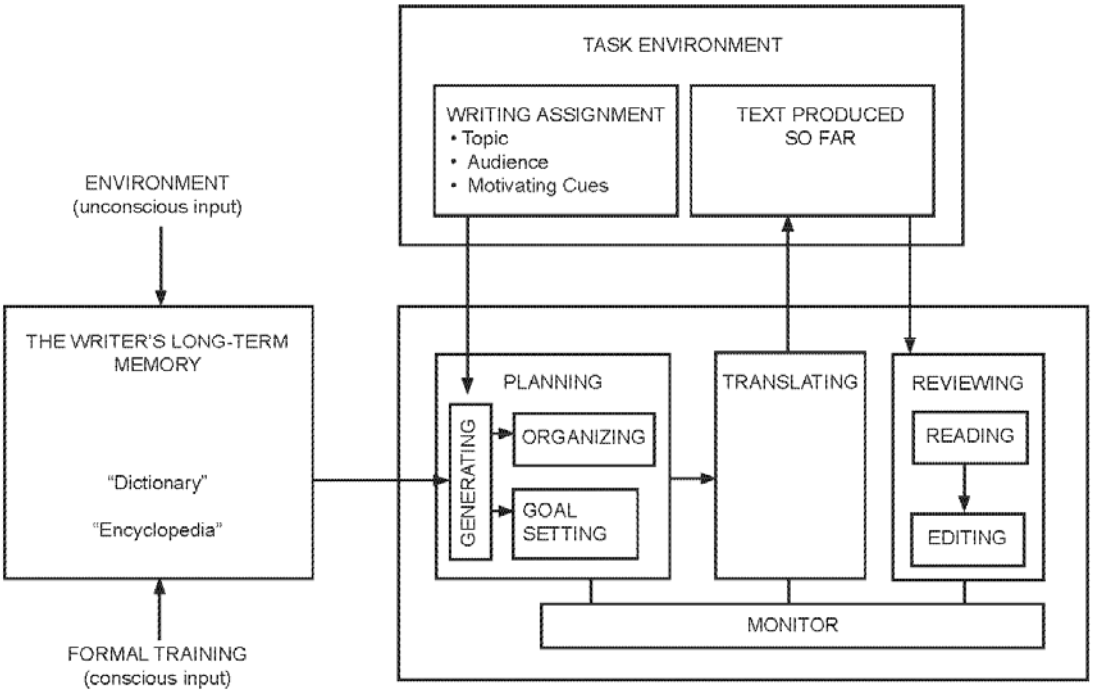


Figure 1. A Modified Version of the Hayes and Flower Writing Model (Ackerman & Smith, 1988, p. 119).

There is general agreement among researchers that the writing process proceeds in three stages from prewriting to writing to revising the finished document (Emig, 1977; Flower & Hayes, 1981; Hillocks, 1984; Lindemann, 2001; Murray, 1968). It also is generally accepted that writers differ in how they approach the process. Although some writers may move linearly through the three stages of the process, most move recursively, adding ideas and content, revising verbal presentation, and altering organizational schemes as the task proceeds. Scholars concerned with the problems of teaching students to write effectively have described the cognitive processes involved in the three stages of this complex task.

Prewriting. Prewriting is considered an important stage in the writing process during which preliminary thinking and planning occur. Individual writers have varied approaches to this stage of writing. Some have personal rituals of preparation. Professional authors typically keep notebooks filled with ideas for later use, while many simply think and plan as they go about daily activities. Even individuals who begin to write immediately at the beginning of a writing session have already carried out mental planning or will incorporate planning as they go. The student who writes an assigned essay in a testing situation, however, must compress this stage of writing to fit given constraints of time, place, and content.

Scholars (Emig, 1977; Flower & Hayes, 1981; Lindemann, 2001; Murray, 1968) describe several prewriting techniques, among them brainstorming and clustering. These two techniques illustrate how a writer begins with a broad idea that eventually leads to a cohesive, well-mapped-out final product. Brainstorming is the act of noting everything that comes to mind on the particular topic; clustering is mapping these ideas into specific groups to determine what information has been generated from brainstorming. A visual representation of clustering ideas and information might assist a writer in organizing thoughts, choosing the focus of the essay, and

locating weaknesses in research and planning. Figure 2 gives an example of clustering for an essay related to the Florida Comprehensive Assessment Test (FCAT). Other prewriting techniques are note taking, outlining, mapping, and scribbling.

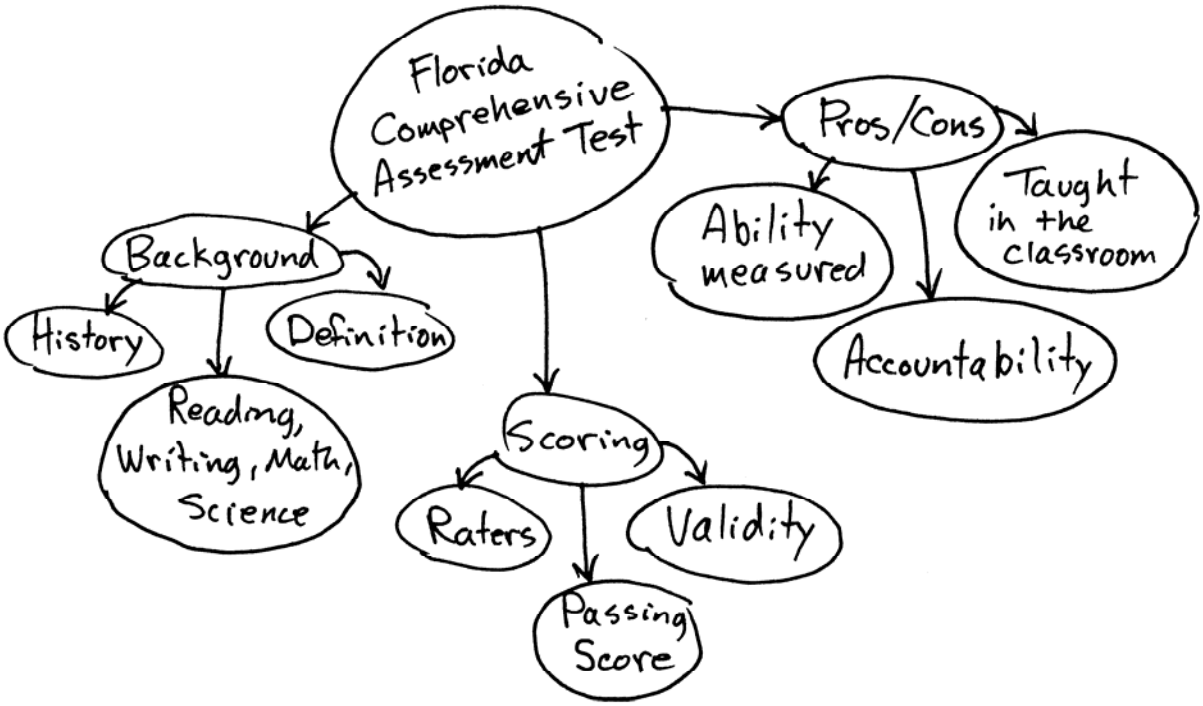


Figure 2. Clustering Related to FCAT Examination Essay Question

Some authorities (Elbow, 1981; Macrorie, 1985; Murray, 1968; Lindemann, 2001) believe that freewriting, which is writing for approximately 10 minutes without concern for grammar, coherence, or sense, will help a writer bypass the critical mind and allow access to fresh ideas that generate greater creativity. Elbow (1981) believes that freewriting as a student activity should occur about three times a week but does not address its appropriateness in a testing situation.

Writing. Although Lindemann (2001) includes writing a first draft as a prewriting activity, the second stage in the writing process is the actual drafting of the document. Most

scholars (Murray, 1968; Flower & Hayes, 1981) believe that the writing stage is simply translating into a visible language the ideas that were generated from planning and organizing. In this stage, the writer determines the focus of the essay, records available knowledge, and identifies areas needing further research. Flower and Hayes (1981) declare that it is in this stage that writers need to be consciously aware of grammar and syntax; most other scholars believe that grammar and editing come last.

Lindemann (2001) believes that this is the stage where writers become cognitively aware of audience, whereas Murray (1968) states that audience is considered in the prewriting stage. Some scholars, such as Emig (1977), believe that audience awareness occurs in the revision stage.

Revising. Scholars agree that revision is required for a successful final product, but they differ on definitions of the process. Murray (1968) believes that revision is rethinking, researching, and rewriting and agrees with Emig (1977) and Flower and Hayes (1981) that revision is continuous throughout the writing process. Lindemann (2001) adds editing to the revision stage, whereas other scholars (Emig, 1977; Hillocks, 1984; Murray, 1968) believe it to be necessary but do not formally include it as part of the writing process.

Sommers (1980) says that unskilled writers define revision as rewording, which refers to cleaning up speech, crossing out what is not needed, and looking for editing mistakes. Hillocks (1984), who conducted a meta-analysis on student writing, found that writers who focused only on grammatical errors during the revision process never improved their ability to write. This may explain why Flower and Hayes (1981) believe that grammatical errors need to be addressed earlier in the writing process and why others believe that editing is only a complement to the revision process.

Link between psychological constructs and composition modes. Lee (2002) attempted to find a relationship between the composition modes and the psychological constructs of students when producing an essay, paralleling Flower and Hayes's (1981) model. Lee posed two questions: (a) What differences exist in composing processes across paper and computer modes? and (b) To what extent and in what ways does the quality of the written products differ across modes?

The participants in Lee's study included six Korean undergraduate and graduate students across disciplines at the University of California, Los Angeles—two in an intermediate English as a Second Language (ESL) class, two in an advanced ESL class, and two completing Ph.D. coursework. The study was completed over a two-day period. The first day, a preliminary questionnaire was distributed that focused on the participants' writing behavior and keyboarding skills. On the second day, the participants took test 1 and then test 2 using the opposite mode. A posttest interview was given after the second test, where the students reported on their composing processes. The procedures led to the following data: qualitative descriptions of the participants' composing processes, average pause time, amount of prewriting, participants' perceptions of the compositions, and scores on analytic rating scales. The testing was videotaped and the interviews were audiotaped.

The results clearly identified two specific areas where the participants' psychological constructs differed across modes related to their use of prewriting and familiarity with the keyboard. Results indicated that participants did more prewriting when composing on paper and pencil. The average time of prewriting for the computer was 81.5 seconds, compared with paper and pencil, which was 226 seconds. The participants' background knowledge across modes affected the way they approached the composing process. Participants who were more familiar

with the keyboard tended to type with no pauses; conversely, participants unfamiliar with the keyboard took frequent pauses.

Five of the six students received higher scores when composing with paper and pencil. The differences in scores between the two modes ranged from 1 point to 3.2 points, which was not significant. One could hypothesize that prewriting is a necessary tool for composing, especially by hand, and has a positive effect on the essay's organizational structure. Lee (2002) concluded that writing on computers forces a new way of thinking about the composing process. Furthermore, Lee argued that the contrast between the modes of composition may lead to differences in test performance.

Relationships Among Student and Rater Characteristics and Composition Mode on Holistic Essay Scores

Powers, Fowles, Farnum, and Ramsey (1994) investigated the effects of mode (handwritten or computer produced) on essay scores of 32 beginning teachers, each of whom produced an essay in each mode. These teachers were selected from a group of more than 500 who had chosen to produce an essay in each mode. All of the original group of papers were scored holistically by two raters on a six-point scale. Handwritten essays of the 32 selected teachers were word processed and their computer-produced essays were converted to handwritten versions. All of these original and converted papers were rescored by two trained readers who were not involved in the original scoring process.

The first part of the study revealed that when scored in their original forms, the handwritten essays (mean = 3.63) received a statistically significant higher mean rating than the computer-produced essays (mean = 3.27). However, the computer-produced essays that were converted to

handwriting yielded a mean of 3.45 while those converted from handwriting to computer-produced yielded a mean of 2.98.

The same data were used in the second part of the study, but four new raters were employed and their training was modified based on possible reasons for the discrepancy between the means of the two modes. The major modifications involved

1. emphasizing that handwritten and word-processed essays may make different impressions. The results of Study 1 were discussed, and readers were encouraged to try to get beyond the mode in which an essay was presented. Staff made it clear, during the training, that [the researchers] were trying to eliminate this effect.
2. discussing the influence of (perceived) length on essay scoring,
3. using both handwritten and word-processed essays in the training, and
4. checking for differences in the standards applied to scoring essays in the two modes. Also, all word-processed essays were double-spaced for this study, so that these versions would not appear to be dramatically shorter than handwritten essays. (Powers et al., 1994, p. 227)

The results of the second part of the study showed that, when the handwritten essays were converted to computer mode, the mean score for the 32 originally handwritten essays decreased from 3.48 to 3.20; for the essays converted from computer mode to handwritten, the mean score increased from 3.48 to 3.75. This reduction of approximately 25% in the discrepancy between ratings of essays written in the two modes over ratings found in the first part was attributed to the modifications of rater training for the second scoring.

Russell and Haney (1997) studied the effect of mode of administration (handwritten versus computer) on multiple-choice and holistic essay scores of middle school students. Students were

enrolled in the Advanced Learning Laboratory School, a school program in Worcester, Massachusetts. Each classroom contained several networked computers that allowed students to do research on the Internet and write reports and papers. In an evaluation conducted prior to the present study, it was found that the quality of open-ended assessments of writing skills had declined. For example, the percentage of satisfactory responses to a long answer item decreased from 71% in the fall of 1993 to 41% in the spring of 1994.

For the study, Russell and Haney (1997) assembled assessment measures consisting of (a) a set of 14 open-ended items (2 writing items, 5 science items, 5 mathematics items, and 2 reading items), (b) 15 National Assessment of Educational Progress (NAEP) language arts items, (c) 23 NAEP science items, (d) 18 NAEP mathematics items, and (e) a writing performance item (PWAvg), which consisted of three holistic reader scorings that were averaged to provide one score. One math, 2 language arts, and 3 science NAEP items were open-ended; the remainder were multiple choice. Experimental and control students (46 versus 68) were selected from those of a larger longitudinal study.

Computer versions of the NAEP items and the writing performance item were prepared for administration to the experimental group. All students were administered the open-ended measure by the handwritten mode.

Tests of statistical significance between the means of the experimental and control groups were made using t-tests. NAEP science and PWAvg were the only measures that produced significant effects. Means and standard deviations for the experimental (computer) and control (handwriting) groups for PWAvg and science are shown in Table 1.

Table 1

Results of NAEP Writing and Performance Items

PWAvg				
Experimental		Control		
Mean	SD	Mean	SD	Effect Size (df)
2.81	0.59	2.30	0.55	0.94
Science				
Experimental		Control		
Mean	SD	Mean	SD	Effect Size (df)
10.68	4.39	8.67	4.17	0.48

The standardized effect size (df) was 0.94 for PWAvg and 0.48 for science. Russell and Haney (1997) reached the following conclusion:

> Our results, if generalizable, suggest that for students accustomed to writing on computer for only a year or two, such estimates of student writing abilities based on responses written by hand may be substantial underestimates of their abilities to write when using a computer.
>
> This suggests that we should exercise considerable caution in making inferences about student abilities based on paper-and-pencil/handwritten tests as students gain more familiarity with writing via computer. And more generally, it suggests an important lesson about test validity. Validity of assessment needs to be considered not simply with respect to the content of instruction, but also with respect to the medium of instruction. As more and more students in schools and colleges do their work with spreadsheets and

> word processors, the traditional paper-and-pencil modes of assessment may fail to measure what they have learned. (p.17)

Russell (1999) conducted a follow-up study to the one by Russell and Haney (1997). The study showed large administration mode effects for handwritten versus computer-produced open-ended science item essays. Subjects were 287 students from two Worcester middle schools, the previously mentioned Advanced Learning Laboratory School and the Sullivan Middle School. Students were randomly assigned to various treatment groups.

The dependent variables were open-ended tests in language arts I (three items), language arts II (three items), mathematics (six items), and science (six items). Independent variables included a keyboarding test, a prior ability test (SAT 9 administered the previous year including reading, mathematics, science, and composite Normal Curve Equivalent scores), and a student questionnaire that asked about the student's computer experience. The keyboarding test contained two passages taken from encyclopedia articles. Students were given two minutes to copy each passage, and words per minute unadjusted for errors were averaged across the two passages as a measure of keyboarding speed. The questionnaire asked

1. how long the student has had a computer in his/her home;
2. how many years they have used a computer;
3. how often they currently use a computer in school or at home;
4. how often they use a computer during different stages of their writing process (e.g., brainstorming, outlining, composing a first draft, editing, writing the final draft); and
5. whether they prefer to write papers on paper or on computer. (Russell, 1999, p. 6)

The open-ended science test was the only instrument for which a statistically significant group difference was found when the level of significance was adjusted for multiple comparisons.

The standardized effect size was .57. The computer group mean was about one-half standard deviation greater than that of the handwriting group. Subgroup analyses showed that, for students whose keyboarding scores were greater than one-half standard deviation above the mean, language arts scores for the computer group were higher than those for the handwritten group. The converse was true for students with keyboarding scores below one-half standard deviation of the mean; the handwritten mode produced higher language arts scores than the computer mode.

Russell and Tao (2004a) partially replicated the findings of Powers et al. (1994) in a study with two parts. The first part sought to determine whether there were composition mode effects (handwritten versus computer written) on the quality of holistically scored essays. The second part was designed to determine whether composition mode affected the visibility of mechanical or structural errors in student-produced essays. Raters were asked to identify the following types of errors:

- spelling,
- punctuation,
- capitalization,
- awkward transitions, and
- confusing phrases or sentences.

The subjects for the study were part of a larger study of 4th-, 8th- and 10th-grade students who took the Minnesota Comprehensive Assessment Systems Language Arts Test in 1999. There were 60 students each in grades 8 and 10 and 52 in grade 4. Originally, all essays were handwritten and transcribed to computer. (All spelling, punctuation, and grammatical errors were included in the transcriptions.) All essay copies were presented to raters in three ways:

1. handwritten,

2. single-spaced 12 point Times Roman computer text, and
3. double-spaced 14 point Times Roman computer text.

For all three grades, mean scores were greater for handwritten responses than for computer-generated transcriptions. In grades 4 and 8, spacing had no effect, but in grade 10, double-spaced responses produced lower means. When the level of significance was adjusted for multiple comparisons, the computer mode produced higher means for spelling errors and confusing passages.

Russell and Tao (2004b) explored causes of the presentation effect (the tendency for raters to give higher ratings to handwritten essays than to computer-produced essays) and ways to reduce or eliminate it. Sixty handwritten essays from eighth-grade students and 12 raters were employed in the study. All 12 raters received three hours of initial training in scoring procedures. All 60 papers were transcribed into computer text and formatted in three ways: (a) single-spaced 12 point Times New Roman font, (b) single-spaced 14 point Script font (to be similar to handwritten cursive writing), and (c) single-spaced 12 point Times New Roman font with all spelling errors corrected. In part one of the study, 4 pairs of raters scored essays from the computer-formatted set plus the handwritten set. None of the raters scored the same essay twice. Two scores were obtained for each essay—topic/idea development and Standard English conventions.

In the second part of the study, the remaining four raters received supplemental training that focused on issues that might influence their scoring procedures. The additional training included

1. reviewing past research on the topic;
2. examining a set of responses from a previous study to compare differences in the apparent lengths of the same responses presented in handwritten and computer-print forms and examining differences in the visibility of spelling, punctuation, and paragraphing errors;

3. scoring a sample of four responses presented in both formats and discussing differences in scores with a specific focus on the influence of appearance;
4. suggesting that raters maintain a mental count of the number of mechanical errors they observe while carefully reading a response; and
5. encouraging raters to think carefully about the factors that influence their judgments before assigning a final score. (Russell & Tao, 2004b, p. 6)

Results of the study indicated that essay format exerted a significant impact on the holistic scores: both scripted computer text and the handwritten responses yielded significantly higher scores than those from computer text. Responses from scripted text did not differ from handwritten responses. When the level of significance was adjusted for multiple comparisons, none of the effects involving spelling corrections were statistically significant.

When raters received supplemental training, there were no statistically significant differences found for any effect. Training raised scores from computer text responses to the same level as scores given to handwritten responses. Russell and Tao (2004b) concluded that ". . . this study provides preliminary evidence that the presentation effect can be eliminated through training" (p. 14).

Wolfe, Bolton, Feltovich, and Welch (1993) conducted a two-part investigation of presentation mode. In the first part, they studied differences between handwritten and word-processed essays given to students in two test administration modes on two different days. The subjects were 157 tenth-grade students from three Midwestern high schools. On the first day, 80 students wrote essay drafts with paper and pen and 77 composed essay drafts on word processors. On the second day, students were asked to review and revise their drafts and to produce a final essay. On this day, they were allowed to choose the mode of administration.

The second part of the study involved determining whether there were differences in the method used to score the handwritten and word-processed essays. All essays in each mode were transcribed to the other mode. Original handwritten papers were transcribed using a variety of font sizes and print qualities, and all original word-processed papers were transcribed to handwritten copies of varying handwritten quality by a number of writers. Each essay was assigned a holistic rating by 2 of 18 independent, experienced raters selected at random.

Results of an analysis of variance showed significant differences between essays scored in word-processed form (3.91) and essays scored in handwritten form (3.39), as well as between essays scored in their original form (3.80) and in the transcribed presentation mode (3.55).

Harrington, Shermis, and Rollins (2000) used 480 students from a large Midwestern university to study differences in the writing quality of essays written by hand and written with word processors. Essays are part of regularly administered English placement exams, and raters use a 22-point holistic scale to place students in appropriate courses. Scores of 1–4 indicate placement in a pre-basic course, 5–11 indicate a basic course assignment, 12–18 indicate a first-year course assignment, and 19–22 indicate honors class assignment.

Test raters' scores indicate whether an exam is prototypical for a given course or on the borderline between two courses. If the first rater places the student in first-year composition, the process is complete; if the first rater places the student in another course, a second (and third, if needed) rater scores the essay.

The students were randomly assigned to one of three groups. The first group wrote their essays in traditional bluebooks. In group two, students wrote in bluebooks but their essays were transcribed by project staff and stored for scoring in a computer database. Spelling and

grammatical errors were not changed during the transcription. In the third group, students used word processors to type their essays directly into the database for scoring.

The placement scores were analyzed using a one-way analysis of variance. Means and standard deviations for the three groups are seen in Table 2.

Table 2

Comparison of Handwritten and Word-Processed Essay Placement Scores

Group	Mean	Standard Deviation
Handwritten	13.14	2.93
Transcribed	12.58	3.06
Word processed	12.54	3.32

The f-ratio was 1.2, a statistically nonsignificant result that indicated no differences in writing quality among the three groups. Inter-rater reliability (the average correlation between pairs of raters) was .87.

Manalo and Wolfe (2000) conducted a study in which the subjects were 152,951 Test of English as a Foreign Language (TOEFL) examinees from 223 countries. The subjects were allowed to choose the administration mode; 51.5% chose the word processor and 48.5% chose paper and pencil. The dependent variable in the study was a holistic score that was the average of scores from two trained independent readers. A control variable—the sum of three multiple-choice TOEFL scores (reading, listening, and structure)—was constructed for use as a covariate. Analysis of covariance was used to analyze the data. The mode effect revealed a statistically significant difference with an effect size of .30 in favor of the handwritten mode.

Manalo and Wolfe (2000) hypothesized that the difference in modes was due to a double translation effect:

> . . . an examinee must first translate the composition from the native language to the English language. Next, the examinee must translate the composition from the English language to the computer language. Hence, the examinee is required to make two cognitive translations—one between the native language and the English language and another between the English language and the computer language (p. 9).

Wolfe and Manalo (2004) studied the effect of composition mode on holistic scores of responses on the writing section of the TOEFL. Subjects were 133,906 individuals who took the TOEFL between January 1998 and February 1999. Fifty-four percent were male, and 46% were female. Examinees were from 200 countries and represented 111 different languages; ages ranged from 15 to 55 years, with an average age of 24.26. Only participants who provided complete TOEFL scores and demographic data were included in the study.

Three sections of the TOEFL consist of multiple-choice items: listening (30 items), structure (20 items), and reading (44 items). The fourth section is a writing test. Scaled listening, structure, and reading scores were averaged to form a composite score—English. The multiple-choice items were computer administered but examinees were given a choice of taking the writing portion in handwritten or word-processing mode. Demographic variables were gender (0 = female, 1 = male) and region of origin, treated nominally (North America, Africa, Asia and Pacific Islands, Central and South America, Europe, and Middle East). Keyboarding was coded as a dichotomous variable depending on whether the examinee's language uses a keyboard containing an alphabet similar to English. (Roman or Cyrillic were coded as 1; others were coded as 0.) The dependent variable in the study was the essay, the average of scores assigned by two raters.

General linear modeling was used in analyzing the data. The results showed a clear difference in means for the English variable for the two essay composition mode groups

(handwritten = 16.68, word processing = 18.60) but only a small difference for essay means (handwritten = 4.06, word processing = 4.09). Because of the large sample size, all of the covariates were statistically significant but not of practical significance.

Breland, Lee, and Muraki (2005) studied TOEFL essays produced between July 1998 and August 2000. The total number of essays available for this time period was 622,588. Eighty-seven different prompts served as topics. Four of the prompts had insufficient data and were eliminated.

About 58.7% of the examinees (365,683) chose to write an essay using a word processor and 41.3% chose the handwritten mode. The examinees represented more than 140 native language groups, with Spanish (10.8%), Arabic (8.8%), Chinese (8.4%), Japanese (7.1%), Korean (6.0%), French (4.5%), and German (4.4%) accounting for half of the population.

The TOEFL computer-based test contains reading, listening, structure, and writing sections, each of which yields a subscore. Reading, listening, and structure scores were combined to create a variable called English Language Ability (ELA), which was used as a control variable. Holistic scores were the average ratings of two raters who were trained using intermingled handwritten and word-processed essays. Logistic regression analysis was used to analyze the data. Table 3 shows the sample sizes, residual means, standard deviations, and effect sizes for essay and ELA scores.

Table 3

TOEFL Essay and English Language Ability Scores

	n	*M*	*SD*	*d*
TOEFL essay scores				
Word-processed group	365,683	4.08	1.03	0.09*
Handwritten group	257,176	3.99	0.89	
English language ability				
Word-processed group	365,683	0.52	2.59	0.47*
Handwritten group	257,176	-0.73	2.72	

* $p < .0001$, two-tailed

Treatment effects for both modes and the ELA control variable showed a statistically significant difference. The mode effect, however, was so small as to be negligible.

Bangert-Drowns (1993) performed a meta-analysis comparing essay scores of students who wrote essays by hand or by using a word processor. The following four criteria were used in selecting 32 reports from approximately 200 studies identified in the Educational Resources Information Center, Dissertation Abstracts, and bibliographies of collected reviews:

> First, each study compared two groups of students who received virtually identical instruction in writing, differing only in that one group was allowed to use the word processor for some phase of the writing process. In all cases, the comparison group wrote by hand. . . . Second, the studies were retrievable from university and college libraries through interlibrary loan, from the Educational Resources Information Center (ERIC), or from University Microfilms International. Third, these reports measured treatment outcomes quantitatively.

> Not all the included studies had sufficient information to permit the calculation of effect sizes, but all provided at least enough information for coding the statistical significance and direction of the results. Fourth, the reports showed no severe methodological flaws. (Bangert-Drowns, 1993, pp. 74–75).

One of the studies contained separate analyses of students with low and middle writing ability, and these analyses were counted as two studies, yielding 33 studies in the sample. The dependent variable of major interest was writing quality as measured by treatment effect sizes (treatment group mean minus the control group mean divided by the pooled standard deviation). Effect sizes could be computed for only 20 of the original 33. The average effect size was 0.27 (SE = 0.11, p = .02), thus indicating a weak significant difference in favor of the word-processing group.

Bangert-Drowns (1993) also studied differences between word-processing and paper-and-pencil administration modes for 21 study features. These features were categorized by instructional treatment, research methodology, study setting, and publication features. Only one feature, writing ability, yielded a statistically significant relationship with effect size. This dichotomous variable was coded 1 if the study provided remedial writing instruction for students who had difficulty with writing and was coded 0 otherwise. Nine of the 20 studies provided remedial instruction. The average effect size, in favor of the word-processing group yielded by these 9 studies, was 0.49, p = .06.

Bridgeman and Cooper (1998) investigated presentation mode differences between handwritten and word-processed holistic essay scores from the Graduate Management Admissions Test (GMAT). A random sample of students who were registered to take the handwritten version of the GMAT in October 1996 were invited to also take a free computerized

version of the test and were told that their computer scores would replace their handwritten scores only if they were higher. Data were obtained from 3,470 examinees of whom 2,453 were from four major ethnic groups (African American, Asian American, Hispanic, and White).

Both versions of the test contained two 30-minute essay questions. One required the examinee to write an analysis of an issue and the other gave an argument and required the examinee to write an analysis of the argument. The computer randomly selected one topic of each type for each person. Order was counterbalanced such that an issue essay was first for half of the sample and an argument essay was first for the other half. For handwritten essays that were part of the regular GMAT October administration, there was only one argument topic and two issue topics (one for the Americas and one for the rest of the world). All students responded to the argument topic first.

Both handwritten essay totals (issue plus argument) and word-processed totals were computed by adding the two scores together. Word-processed scores were subtracted from the handwritten ones to produce a different score with positive values showing better performance on the handwritten essays. All statistical analyses favored the handwritten essays with only word-processing experience having a statistically significant effect. Rater reliability was estimated by using the Spearman-Brown formula. The estimate was higher for word-processed essays (.87) than for handwritten ones (.80).

Summary

Psychological constructs associated with writing are probably more closely related to the quality of a completed document than to the cognitive processes involved in producing it. Although authorities in the field differ in their descriptions of these processes, they agree that writing involves three stages: deciding what is going to be said (prewriting), fashioning the

selected content into a coherent whole (writing), and reviewing/revising the document to ensure that it presents the content in the way the writer intends and is as error free as possible (revising). It is unknown whether there are identifiable links between mode of production or essay quality and any one or more of the specific cognitive processes involved in essay composition.

There appears to be a weak, if any, relationship between mode of production (handwritten versus computer produced) and the quality of student essays. Four of the 12 studies reviewed here showed no significant difference between modes, 3 yielded results favoring the quality of computer-produced essays, and 5 showed higher average scores for handwritten essays. Ages and native languages of subjects (English or other) had no apparent effects on results nor did the time period during which the studies were conducted (1990–1999, 2000 and beyond).

It is important to note, however, that production mode appears to cause the scores to be higher for handwritten essays than for word-processed ones. This is true even when handwritten essays are transcribed and each set of essays is rated by different raters. Because rater bias is the most probable cause of this result, it is imperative that rater training includes working with essays from both modes, identifying specific mode characteristics (e.g., neatness, legibility, apparent length) so raters will minimize their influence on writing quality scores.

Methodology

Subjects

Subjects were undergraduate and graduate students in the teacher training programs of Florida's colleges and universities. The total number of subjects was 66 students. Eleven students dropped out of the study after they finished their paper-and-pencil test session. The participants, who represented two major Florida regions (north and south), were from Florida

State University (FSU), University of North Florida (UNF), Miami-Dade College InterAmerican campus, and Chipola College. For the central region of Florida, efforts to recruit participants and secure permission from the human subjects review committees were unsuccessful.

Instruments

The primary measure used in this study was the essay portion of the FTCE. For this purpose, the FDOE released four expository essay prompts, the corresponding FTCE scoring six-point rubric, the Essay Topic Booklet, and the Written Response Booklet. FDOE also provided the score distributions and means from prior administrations for two of the prompts. Other instruments included

1. a questionnaire designed by FSU/CALA and administered to the students that contained items related to computer familiarity and use, and attitudes toward the two modes of administration (Appendix A);
2. a similar questionnaire developed and administered to the holistic scorers (Appendix B);
3. a record of student prewriting activities obtained from each participant for both modes of writing; and
4. a demographic form that captured information during the computer-based administration.

Procedures

The subjects were asked to take an essay test in each of two consecutive sessions, one using paper and pencil and one using a computer. For the UNF administration, students were tested on the two modes of writing during the same day, with a two-hour break in between. Miami-Dade College and Chipola College administrations were on consecutive days. For the FSU administration, students were tested on one mode on a Monday and the next mode on a Wednesday. With the exception of UNF, students used the paper-and-pencil mode for the first

session and the computer-based mode for the second. The order was reversed for UNF because of computer lab availability. For each administration, the students could select from one of two prompts. The computer-based administrations used prompts 1 and 2, and the paper-and-pencil administrations used prompts 3 and 4.

For both modes of writing, students were assigned examinee ID numbers that consisted of an institutional identifier and a three-digit number. Prior to administration of the examinations, each participant signed a consent form.

Standard writing tools (pens and pencils), an Essay Topic Booklet (containing the topics and space to plan the essay), and a Written Response Booklet (where the final essay is written) were given to the subjects for the handwriting mode of the essay. The Essay Topic Booklet and Written Response Booklet were close replications of the booklets used by FDOE to administer the FTCE Essay examination. Test sessions for handwriting administrations were approximately 60–75 minutes long, with 10–15 minutes for general instruction and 50 minutes for prewriting and writing activities.

Students used an Internet-based word-processing program to write their computer-based essays. The program provided a preplanning section on the same screen with the essay writing section. Students also were given an Essay Topic Booklet to use for preplanning if they preferred to plan their essay on paper instead of using the computer. Formatting was preset to ensure margins, fonts, and spacing were consistent, and students were prevented from using tools for grammar or spelling corrections. Test sessions for computer-based administrations were approximately 70–80 minutes long, with 10–15 minutes for general instruction, 50 minutes for prewriting and writing activities, and 10–15 minutes for completing a questionnaire.

All of the papers, excluding the anchor papers, were transcribed (see Figure 3). Anchor papers are the essays chosen from those written to be representative samples for each point on the six-point rubric. The paper-and-pencil essays were arranged numerically by examinee ID number and every other paper was selected for transcription to word-processed format. The remaining ID numbers were used to determine the computer-based papers to be transcribed to handwritten format. Twenty-two essays were transcribed from handwritten to computer, and 23 were transcribed from computer to handwritten. During the process of transcription from handwritten to computer, care was taken to capture the students' essays exactly, including errors, so that raters would not be able to discern the differences. The same procedure was used for transcribing from computer to handwritten essay. Multiple transcribers were used to produce a variety of handwriting styles.

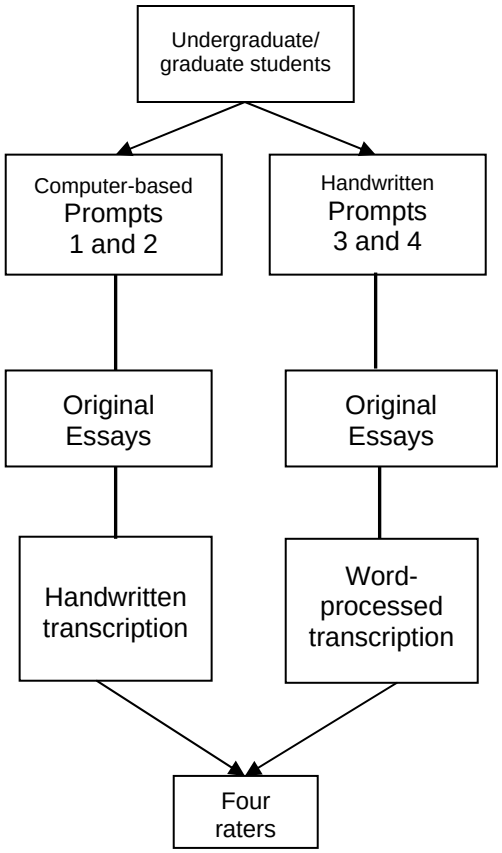


Figure 3. Essay Administration/ Scoring Overview

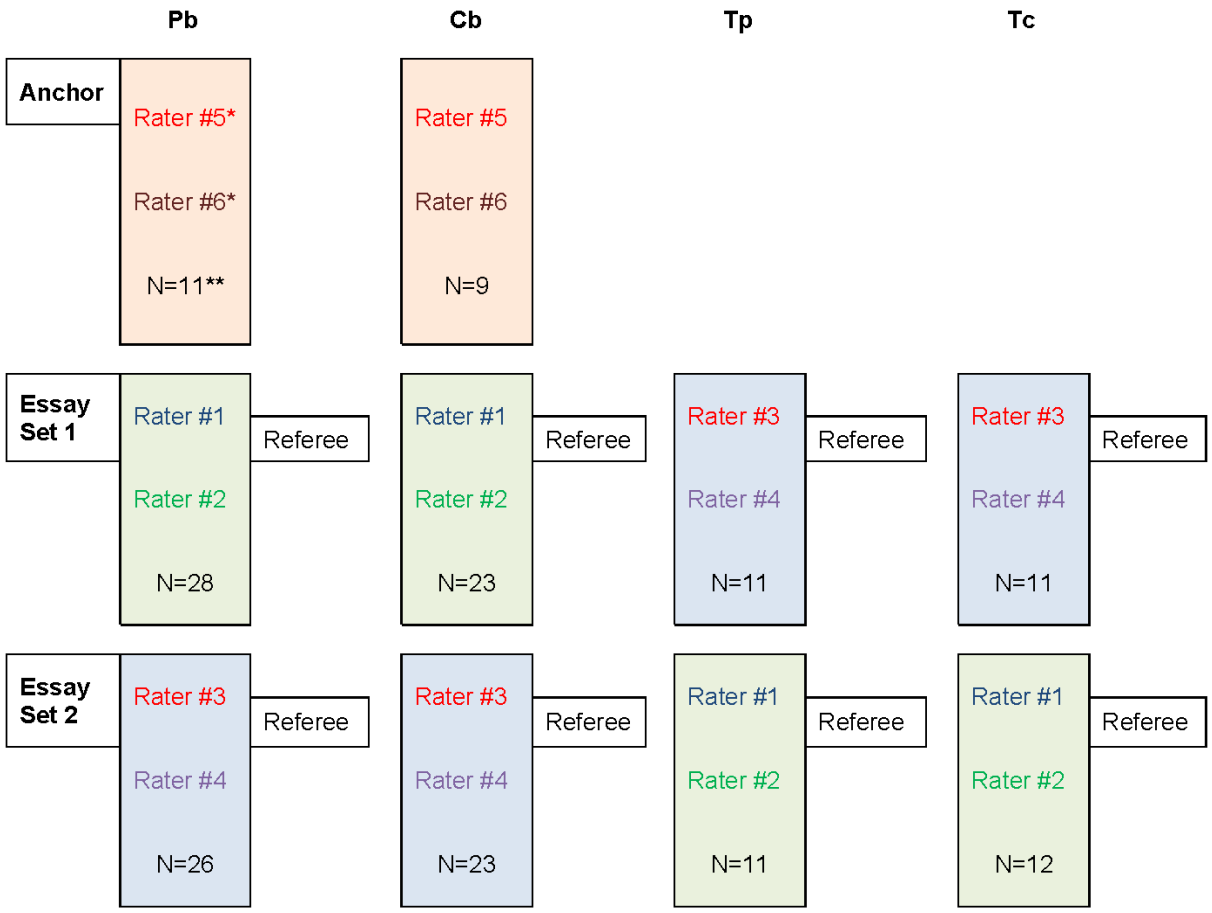
Because anchor papers were not available from FDOE for the four prompts, two experts in holistic scoring reviewed, scored, selected, and annotated subjects' essays to represent each point on the six-point rubric. One of these experts also helped set up the scoring procedure, select four raters with previous holistic scoring experience, and conduct holistic training for the raters. Essays were selected to be used for discussion during the training session. The selections included 11 essays from the paper-and-pencil examination and 9 from the computer-based

examination. Modifications in training were made based on findings from previous research, expanding these procedures used by Powers et al. (1994):

1. emphasizing that handwritten and word-processed essays may make different impressions … readers were encouraged to try to get beyond the mode in which an essay was presented. Staff made it clear, during the training, that [the researchers] were trying to eliminate this effect.
2. discussing the influence of (perceived) length on essay scoring,
3. using both handwritten and word-processed essays in the training, and
4. checking for differences in the standards applied to scoring essays in the two modes. (p. 227)

Scoring

Each essay was scored by two raters using the six-point rubric. Accompanying each essay was a scoring form used to capture the topic ID, preplanning status, rater IDs (for raters 1–4), assigned scores, referee status, and total score (Appendix C). To keep scores confidential, raters were assigned individual scoring codes. When scoring discrepancies of more than one point occurred, a referee read the essay and replaced one of the raters' scores. The total score combined both raters' scores and ranged from 0 (off topic) to 12 (maximum score). Each team of raters was assigned the same set of students for both paper-and-pencil and computer-based essays. Raters 5 and 6 were the two holistic experts who selected the anchor and discussion essays (see Figure 4).



Pb = paper-based essay; Cb = computer-based essay; Tp = essays transcribed from paper to computer; Tc = essays transcribed from computer to paper.
*Raters 5 and 6 were the holistic scoring experts.
**Eleven participants took only the paper-and-pencil exams; of these essays, 10 were selected as anchors.

Figure 4. Holistic Scoring Diagram

Data Analysis

A computer data file was created that combined the demographic data gathered through the computer-based administration and scoring forms. The file contained the following information from the subjects:

- examinee ID number,
- gender,
- age,

- race ("Do not wish to disclose" was an option),
- primary language, and
- responses to items on the student questionnaire.

For each subject's essay in each production mode, these data were also contained in the file:

- prompt chosen,
- whether preplanning had been done,
- first rater's ID number,
- holistic score assigned by first rater,
- second rater's ID number,
- holistic score assigned by second rater,
- whether essay had been refereed,
- whether essay was off-topic, and
- total holistic score (the sum of the two raters' assigned scores).

The data were analyzed using both qualitative and quantitative methods. Frequency statistics were compiled for all student demographics and student responses to questionnaire items.

Statistical comparisons of means of transcribed and originally produced papers were made to separate influences of mode of writing. Total scores were used in paired comparison t-tests to investigate the relationships between ratings for paper-based essays (Pb) and those essays transcribed from paper to computer (Tp), and computer-based essays (Cb) and those essays transcribed from computer to paper (Tc). Comparisons were also made between each rater's scores for the originally produced paper-and-pencil and computer-based essays for the same

student. Both parametric (t-test) and nonparametric (Wilcoxon Signed Ranks Test) tests were applied to these comparisons.

The open-ended questions from the student and rater questionnaires were analyzed using qualitative methods. The students' preplanning responses from paper-based and computer-based administrations were analyzed through inspection.

Results and Discussion

Subjects

The total number of participants who wrote essays in both paper-and-pencil and computer-based modes was 55. Eleven other participants who dropped out of the study after their paper-and-pencil test session were excluded from the study. Twenty-six (47%) of the subjects were enrolled at Chipola College, 13 (24%) at FSU, 11 (20%) at Miami-Dade College InterAmerican campus, and 5 (9%) at UNF. Tables 4–8 show a summary of demographic information.

Table 4

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	40	72.7	72.7	72.7
	Male	15	27.3	27.3	100.0
	Total	55	100.0	100.0	

Table 5

Language

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	English	51	92.7	92.7	92.7
	Other	4	7.3	7.3	100.0
	Total	55	100.0	100.0	

Table 6

Age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	19	7	12.7	12.7	12.7
	20	11	20.0	20.0	32.7
	21	4	7.3	7.3	40.0
	22	1	1.8	1.8	41.8
	23	4	7.3	7.3	49.1
	24	2	3.6	3.6	52.7
	25	6	10.9	10.9	63.6
Subtotal		35	63.6	63.6	63.6
	28	2	3.6	3.6	67.3
	29	2	3.6	3.6	70.9
	30	3	5.5	5.5	76.4
	31	1	1.8	1.8	78.2
	33	1	1.8	1.8	80.0
	34	1	1.8	1.8	81.8
	35	1	1.8	1.8	83.6
Subtotal		11	19.90	19.90	83.6
	38	3	5.5	5.5	89.1
	39	1	1.8	1.8	90.9
	46	1	1.8	1.8	92.7
	47	2	3.6	3.6	96.4
	51	1	1.8	1.8	98.2
	60	1	1.8	1.8	100.0
Subtotal		9	16.30	16.30	100.0
Total		55	100.0	100.0	

Table 7

Education

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Graduate	7	12.7	12.7	12.7
	Undergraduate	48	87.3	87.3	100.0
	Total	55	100.0	100.0	

Table 8

Race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Asian or Pacific Islander	1	1.8	1.8	1.8
	Black non-Hispanic	14	25.5	25.5	27.3
	Do not wish to disclose	4	7.3	7.3	34.6
	Hispanic	5	9.1	9.1	43.7
	Multiracial/ethnic	1	1.8	1.8	45.5
	White non-Hispanic	30	54.5	54.5	100.0
	Total	55	100.0	100.0	

The majority of participants selected prompt 1 for computer-based essays and prompt 4 for paper-based essays. It should be noted that the six holistic scorers deemed the pairs of prompts of equivalent difficulty. Table 9 provides frequency and percentages for each of the four prompts.

Table 9

Prompt Selections

Prompts	Mode of Writing	Frequency	Percent	Cumulative Percent
1	Cb	44	80	
2	Cb	11	20	100
3	Pb	18	33	
4	Pb	37	67	100

Comparisons of the percentage of the participants who had completed preplanning activities on the paper-and-pencil mode versus the computer-based mode indicated a difference of 7%. An interesting discovery was that of 71% who did preplanning activities during the computer-based administration, 49% chose to do their preplanning on paper rather than on the computer. An additional 7% used both paper and pencil and the computer for preplanning during the computer-based administration. Participants in the computer-based administration had the choice of using the Essay Topic Booklets or a preplanning section on the computer for preplanning activities. Table 10 provides a summary of the modes of preplanning activities for the computer-based administration.

Table 10

Computer-based Preplanning Activities

Preplanning	Frequency	Percent	Valid Percent	Cumulative Percent
None	16	29.1	29.1	29.1
On paper	27	49.1	49.1	78.2
On computer	8	14.5	14.5	92.7
On paper and computer	4	7.3	7.3	100.0
Total	55	100.0	100.0	

Essay Quality and Systemic Rater Bias

Scores on a scale of 0–6 were assigned by raters to each essay they evaluated. One student received a score of 0 because of an off-topic essay; the rest of the scores ranged from 2 to 6. Raters 1 and 2 acted as a team to rate a set of papers in each mode and the transcriptions; raters 3 and 4 comprised a similar team for another set of papers. Raters 5 and 6 were holistic scoring experts who selected anchor papers and served as referee raters who evaluated essays only when scores assigned by the regular raters differed by more than 1 point.

Means of total scores for Pb essays were paired with those of Cb essays, Tp essays, and Tc essays for statistical comparison. Means of total scores for Cb essays were also paired for comparison with those of Tp and Tc essays. Table 11 shows the results of these analyses.

Table 11

Paired Comparisons of Essay Mean Total Scores

						Paired Differences							
									95% Confidence Interval of the Difference				
Pairs	Mean	N	Std. Deviation	Std. Error Mean	Correlation	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	Df	Sig. (2-tailed)
Pb Total	8.51	55	1.762	.238	.320	-.036	2.472	.333	-.705	.632	-.109	54	.914
Cb Total	8.55	55	2.387	.322									
Pb Total	8.50	22	1.596	.340	.415	.227	1.798	.383	-.570	1.024	.593	21	.560
Tp Total	8.27	22	1.723	.367									
Cb Total	8.96	23	1.918	.400	.659	-.913	1.564	.326	-1.589	-.237	-2.799	22	.010
Tc Total	9.87	23	1.866	.389									
Pb Total	9.09	23	1.564	.326	.502	-.783	1.731	.361	-1.531	-.034	-2.168	22	.041
Tc Total	9.87	23	1.866	.389									
Cb Total	8.77	22	2.022	.431	.251	.500	2.304	.491	-.522	1.522	1.018	21	.320
Tp Total	8.27	22	1.723	.367									

Paired comparisons of scores from first and second readings of the essays were made for Pb, Cb, Tp, and Tc essays. Results of these analyses are shown in Table 12.

Table 12

Paired Comparisons of Mean Scores for First and Second Readings

						Paired Differences							
									95% Confidence Interval of the Difference				
Pairs	Mean	N	Std. Deviation	Std. Error Mean	Correlation	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	Df	Sig. (2-tailed)
Pb reading 1	4.24	55	.981	.132	.698	-.036	.744	.100	-.238	.165	-.362	54	.719
Pb reading 2	4.27	55	.932	.126									
Cb reading 1	4.29	55	1.242	.168	.824	-.036	.719	.097	-.231	.158	-.375	54	.709
Cb reading 2	4.33	55	1.171	.158									
Tp reading 1	4.23	22	.752	.160	.835	.182	.588	.125	-.079	.443	1.449	21	.162
Tp reading 2	4.05	22	1.046	.223									
Tc reading 1	5.00	23	1.000	.209	.717	.130	.757	.158	-.197	.458	.826	22	.418
Tc reading 2	4.87	23	1.014	.211									

Significance levels shown in Table 11 were corrected by use of the Bonferroni method for corrections for multiple comparisons. Given the 9 comparisons (Tables 11 and 12), a significance level of .0003 is required to reject the null hypothesis. It can be seen, therefore, that none of these differences between means are statistically significant. It follows that this study offers strong support to the hypothesis that there is no difference in quality between handwritten and computer-produced essays. In addition, these results demonstrate that no systemic rater bias was present in this study.

A nonparametric statistic (Wilcoxon Signed Ranks Test) was also applied to the ratings data in order to discover whether the small sample size had an impact on the t-test results. Again, no significant differences were found among means of the pairs that were compared. Table 13 shows the results of this analysis.

Table 13

Paired Comparisons of Raters' Pb and Cb Scores

						Paired Differences							
									95% Confidence Interval of the Difference				
Rater ID	Mean	N	Std. Deviation	Std. Error Mean	Correlation	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	Df	Sig. (2-tailed)
Pb rater #ID 1	4.6250	8	.91613	.32390	.000	-.37500	1.06066	.37500	-1.26173	.51173	-1.000	7	.351
Cb rater #ID 1	5.0000	8	.53452	.18898									
Pb rater #ID 2	4.6250	8	.91613	.32390	.000	-.37500	1.06066	.37500	-1.26173	.51173	-1.000	7	.351
Cb rater #ID 2	5.0000	8	.53452	.18898									
Pb rater #ID 3	4.4286	7	.97590	.36886	.062	.85714	1.21499	.45922	-.26653	1.98082	1.867	6	.111
Cb rater #ID 3	3.5714	7	.78680	.29738									
Pb rater #ID 4	4.4286	7	.97590	.36886	.062	.85714	1.21499	.45922	-.26653	1.98082	1.867	6	.111
Cb rater #ID 4	3.5714	7	.78680	.29738									

It should be noted that the mean for prompts 1 and 2 in this study (8.55) was much higher than the mean reported by FDOE for the same prompts. Backgrounds of the FDOE candidates are unknown, but it is probable that they included individuals who were not prepared by either training or experience to enter the teaching profession. Subjects for this study were, on the other hand, enrolled in academic teacher training programs.

Individual Rater Bias

The paired comparisons of several combinations of total Pb and Cb scores shown in Tables 11, 12, and 13 were used to investigate rater bias as well as differences in quality of essays in both production modes. However, because composite scores, even those combining only the two scores of a two-person team, can easily mask bias on the part of an individual rater, means were also computed for scores assigned by each rater for each production mode. Numbers of essays read by each rater in each mode were slightly different. Use of paired comparisons would have necessitated the omission of some scores for this analysis; therefore, means only were computed and compared by inspection. Table 14 shows for each rater in each mode the number of subjects rated and the mean scores assigned on the original essays.

Table 14

Mean Scores Assigned by Individual Raters to Original Essays

Rater	N	Pb Score	N	Cb Score
1	28	4.11	23	4.87
2	28	4.25	23	4.35
3	26	4.00	23	4.09
4	26	4.11	23	4.43

Results for comparisons between ratings on the transcribed essays are slightly different from those found for the original essays. Table 15 shows the number of subjects and the mean scores for each of the four raters used in this part of the study.

Table 15

Mean Scores Assigned by Individual Raters to Transcribed Essays

Rater	N	Tp Score	N	Tc Score
1	11	4.45	11	5.18
2	11	4.18	11	5.36
3	11	3.91	11	4.66
4	11	4.00	11	4.45

Mean scores for all raters ranged from 3.91 to 5.36. Both of the extremes were found in ratings of the transcribed essays. Although chance distribution of the essays is the most probable explanation of this finding, it is possible that other factors such as essay length and physical appearance of the handwritten papers are related. In any case, all mean differences are nonsignificant. This finding, combined with results of the paired comparisons shown in Tables 12 and 13, strongly supports the conclusion that neither systemic nor individual rater bias was a factor in this study.

Refereed Essays

The data regarding the number of referred essays in each mode of writing also were analyzed to determine if the modes caused the disparity in raters' scores. Table 16 provides the student ID number, the prompt to which the essay was written for each mode, and the rater ID number.

Table 16

Refereed Essays

	Pb		Cb		Tp		Tc	
Student ID #	Prompt	Rater ID #	Prompt	Rater ID #	Prompt	Rater ID #	Prompt	Rater ID #
001							1	3, 4
003	4	3, 4						
006	3	1, 2						
007	4	3, 4						
011	4	3, 4						
012	3	1, 2						
022			2	1, 2				
102	4	1, 2						
212					4	1, 2		
221					4	3, 4		
304	3	1, 2						
314	3	3, 4	2	1, 2				

The analysis indicates that the mode of writing had an impact on the number of essays that required a referee to cast the final score. More essays in paper-and-pencil mode required a third reading than essays written on computer. In addition, none of these handwritten essays required a referee when they were transcribed to the computer. A possible explanation for this discrepancy is that the overall quality of writing may be more transparent in word-processed form without interference by the writer's handwriting quality or corrections.

Student Questionnaire Results

Most participants in the study own a personal computer (51 of 55), and three-fourths of these have owned their computer for more than three years. While very few of the students use a computer to take notes in class (5 of 55), nearly all (54) use a computer to write papers. The majority indicated that their keyboarding skills were "adequate" (20) or "good" (34). More than 80% of the respondents indicated a preference for writing an essay on computer. These data suggest that most examinees felt reasonably comfortable writing on a computer.

When comparing the participants' age with their mode preference for writing essays, more of the participants 38 years of age and older (4 out of 9) preferred the handwriting mode than participants 37 years of age and younger (6 of 46).

More than half of the 55 respondents (30) indicated that they sometimes preplan or create a rough draft on paper when handwriting an essay, while 24 respondents indicated that they sometimes do so when using the computer to write an essay. Ten students indicated that they never preplan or create a rough draft on paper when handwriting an essay; in contrast, 35 students indicated that they never use the computer to preplan or create a draft when writing an essay on computer. This finding appears consistent with students' indications that they find editing, revising, and organizing information much easier on computer than on paper.

Open-Ended Questions

Question 11: If you answered YES to question 10, please indicate which essay you felt was superior. Explain the ways in which the superior paper was better and whether you think the writing mode was a factor in the difference in quality between these two essays.

Although many of the responses to this question focused more on *why* the examinee preferred writing the essay on computer—rather than which *essay* was superior and why—responses that included only positive statements about writing on computer were considered to indicate that the examinee believed the computer-written essay was superior.

Of the 30 students who responded to this question, 24 stated or inferred that their computer-written essay was superior to their handwritten essay. Five examinees indicated a preference for the handwritten essay; however, two of these attributed the higher quality of the

handwritten essay to their interaction with the topic rather than to the mode of writing. One respondent failed to identify which essay was superior.

Three of the five students who indicated that their handwritten essays were better were correct in their assessments. Actual differences in the paper-based and computer-based scores ranged from 1 to 4 points.

One student said that there was a difference in quality between his two essays but did not specify which essay. The difference between his two scores was 7 points, and his written answer to the question was “Ability to edit.” Another student who preferred her paper-based essay received 5 points higher on it than on the computer-based one. Her written answer was “The paper essay seems superior in content.”

The situation was different for the group who said that their computer essays were better than their handwritten ones. Nine of them received lower scores on the computer than on the paper-based essay, 3 received the same score on both essays, and 12 received a higher score for the computer-based essay. Most of the scores differed by only 1 or 2 points.

The mean combined total score (Pb total plus Cb total) earned by students who accurately judged the differences in quality between their two essays was 17. For those who judged these differences inaccurately, the mean combined total score was 14.67.

More than half (14) of the examinees who indicated the quality of the computer-written paper was superior supported their assertion by stating they could revise and edit their writing more easily on the computer. More than one-third (9) of the examinees cited speed as a reason they preferred the computer-written essay to the handwritten one, with comments such as “I type faster than I write, so it was easier to get all of my ideas down” and “The computer helped me think faster.” Seven examinees identified the ability to better organize ideas and information as a

reason they preferred the computer-written essay (e.g., "… I felt it was more organized, more clear, and I knew what I wanted to talk about more" and "… it's easier to rearrange your thoughts …"). Three examinees referred to a neater appearance as one of the reasons that the computer-written essay was superior. One of these examinees believed that scores might be affected by the number of corrections and neatness of the final essay.

Question 12: For this research, did your planning (prewriting, notetaking, drafting) for writing the two essays differ when you used a computer versus pencil and paper? __Yes __No
Explain your answer.

Of the 48 students who responded to this question, 35 indicated that their preplanning did not differ according to mode of writing and 13 responded that their preplanning was different. Eleven of the 26 students who included explanations in their responses indicated that they felt less need for preplanning on the computer because of the ability to revise as they wrote (e.g., "I really don't have to use prewriting when using the computer because you can edit easier."). Ten students responded that their preplanning was the same or nearly the same for both modes of writing (e.g., "I used the same paper method both times."). Five students indicated that they did not preplan for either mode of writing (e.g., "I don't plan when I write.").

Question 13: Did you revise your essays in the same way when using computer versus paper and pencil? __Yes __No Explain your answer.

The number of students responding that they revised their essays differently across the two writing modes (24) was the same as the number who indicated that they revised their essays in the same manner (24). Of the 48 examinees responding to the yes/no component of this

question, 26 provided explanations. More than half of the explanations (14) cited the ease of editing and moving information in their response (e.g., “It is easier for me to reread on computer than it is on paper. It is also a lot easier to correct your mistakes on the computer.”). Five students explained that they revised in much the same manner for both modes of writing (e.g., “I always proofread my essays, no matter where or how I write them.”). The explanations of 7 students did not clearly support their yes/no responses.

Rater Questionnaire Results

All four raters had owned a personal computer for more than three years. They rated their keyboarding skills as “good” and indicated that they usually use a computer to write papers and prepare and/or deliver instruction.

Three of four raters indicated that they sometimes preplan or do a rough draft on paper both when handwriting and when using the computer to produce a document. Two indicated they never preplan or draft on computer when using the computer to write a document, and two indicated that they sometimes do so. Likewise, the group was split in their preference for writing documents on computer versus paper. While raters expressed their appreciation for the “personal connection” of handwriting a document, they acknowledged the efficiency and convenience of composing and writing on computer. Two of the four raters indicated that the mode of writing (computer versus handwriting) had no effect on the quality of documents, while the other two indicated that the quality of computer-written documents is comparable to that of handwritten documents.

In response to an open-ended question about the most effective mode of writing, the raters stated that they preferred the convenience and efficiency of using the computer to write.

However, they also described handwriting as part of the creative process (e.g., “The personal connection from mind and heart to pen and paper creates a strong bond . . .”).

No information in the qualitative data collected from the raters would predict rater bias toward or against one writing mode.

Conclusions and Recommendations

Research findings have varied regarding the effects of writing mode (paper and pencil versus computer) on students’ performance and, subsequently, their scores. Some studies showed no significant differences between the mode of writing and the quality of student performance. Other studies indicated higher quality for handwritten essays, while others showed higher quality for computer-produced essays.

This study determined that the production mode used by FTCE examinees to respond to an essay prompt is not significantly related to the quality of their essays. Neither systemic bias nor individual rater bias was present as a factor in quality judgments of raters. This result might be attributed to the careful selection of raters and the thoroughness of their training.

The analysis of the data from the student and rater questionnaires revealed that the majority in the study owned a personal computer and felt reasonably comfortable writing on a computer. Nearly all participants also indicated that they use a computer for routinely writing papers, with the majority preferring computer-produced essays to handwritten ones. Eighty-two percent of the participants indicated that they preferred to write essays in computer mode, and 55% believed that they produced better essays in that mode.

The preference for using paper and pencil over computer was higher among participants 38 years of age and older, compared with participants 37 years of age and younger. However, the

data related to the preferred writing mode and the actual score did not indicate that those who preferred the handwriting mode scored lower on the computer.

The findings on preplanning showed that a higher percentage of participants skipped the preplanning stage in the computer-based mode than in the handwriting mode. A possible explanation might be that the word-processing capabilities of the computer gave participants the flexibility to move text and revise. This more recursive approach to the composing process thus reduced the necessity for preplanning. One of the most interesting findings of the study was that a significant number of participants in the computer-based administration continued to do their preplanning on paper rather than on the computer.

In conclusion, this study provides empirical support for using the computer-based writing assessment for the FTCE Essay examination. There are a few issues, however, that should be considered prior to full implementation of the computer-based administration. The most important factor is the quality of training offered to the holistic scorers. During the selection of anchor papers and the training of raters, both handwritten and computer-written essays should be used. In addition, the raters should be informed of the research findings about rater bias.

If the computer-based administration is adopted, materials should be provided for preplanning on paper. It also might be advisable to offer examinees a choice between computer-based and paper-and-pencil administration for a few more years so that the perceived comfort level is higher, especially for older examinees.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12*(2), 117–128.

Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research, 63*(1), 69–93.

Breland, H., Lee, Y. W., & Muraki, E. (2005). Comparability of TOEFL CBT essay prompts: Response mode analyses. *Educational and Psychological Measurement, 65*(4), 577–595.

Bridgeman, B., & Cooper, P. (1998). *Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test.* Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Elbow, P. (1981). *Writing with power: Techniques for mastering the writing process*. New York: Oxford University Press.

Emig, J. (1977). Writing as a mode of learning. In V. Villanueva (Ed.), *Cross-talk in comp theory* (pp. 7–16). Urbana, IL: The National Council of Teachers of English.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. In V. Villanueva (Ed.), *Cross-talk in comp theory* (pp. 273–297). Urbana, IL: The National Council of Teachers of English.

Harrington, S., Shermis, M. D., & Rollins, A. D. (2000). The influence of word processing on English placement test results. *Computers and Composition, 17*(2), 197–210.

Hillocks, G., Jr. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies (in reading and writing achievement in the American schools). *American Journal of Education, 93*(1), 133–170.

Lee, Y. (2002). A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing, 8*(2), 135–157.

Lindemann, E. (2001). *A rhetoric for writing teachers* (4th ed.). New York: Oxford University Press.

Macrorie, K. (1985). *Telling writing* (4th ed.). Upper Montclair, NJ: Boynton/Cook.

Manalo, J. R., & Wolfe, E. W. (2000). *A comparison of word-processed and handwritten essays written for the Test of English as a Foreign Language*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Murray, D. (1968). Teach writing as a process not product. In V. Villanueva (Ed.), *Cross-talk in comp theory* (pp. 3–6). Urbana, IL: The National Council of Teachers of English.

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*(3), 220–233.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education and Policy Analysis Archives, 7*(20), 2–46.

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 and Policy Analysis Archives, 5*(3), 1–21.

Russell, M., & Tao, W. (2004a). Effects of handwriting and computer-print on composition scores: A follow-up to Powers, Fowles, Farnum, & Ramsey. *Education and Policy Analysis Archives, 9*(1), 1–12.

Russell, M., & Tao, W. (2004b). The influence of computer print on rater scores. *Practical Assessment, Research & Evaluation, 9*(10), 1–16.

Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. In V. Villanueva (Ed.), *Cross-talk in comp theory* (pp. 43–54). Urbana, IL: The National Council of Teachers of English.

Wolfe, E. W., Bolton, S., Feltovich, B., & Welch, C. (1993). *A comparison of word-processed and handwritten essays from a standardized writing assessment*. Iowa City, IA: American College Testing Program.

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.

Author Note

FJ King, Faranak Rohani, Carol Sanfilippo, and Natalee White, Center for Advancement of Learning and Assessment, Florida State University.

The research described herein was funded by the Florida Department of Education. The findings and opinions do not necessarily reflect the positions or policies of the Florida Assessment and School Performance Office or the Florida Department of Education.

Our special thanks to those who assisted in the research project by facilitating the human subjects review process and arranging for students to participate: Lou Cleveland and Gina McCallister, Chipola College; Pamela Carroll and Joseph Valente, Florida State University; M. Victoria Florit, Miami-Dade College InterAmerican Campus; and Lunetta Williams and Nicole Sayers, University of North Florida. Thanks to those who offered their expertise for the holistic scoring process: Linda Clarke, Larry Crombie, Ruth Garrison, Wolfgang Lepschy, Elizabeth Novinger, and Lorrie O'Dell.

Appendix A

Student Questionnaire

Examinee ID Number _____

1. Do you own a personal computer?
   ❑ Y ❑ N

2. If yes, how long have you owned a computer?
   ❑ less than a year
   ❑ 1–3 years
   ❑ more than 3 years

3. Do you usually use a personal computer for writing papers?
   ❑ Y ❑ N

4. Do you use a computer to take notes in classrooms?
   ❑ Y ❑ N

5. How would you rate your keyboarding skills?
   ❑ poor
   ❑ just adequate
   ❑ good

6. When you are writing an essay on **paper,** do you do preplanning or a rough draft on **paper** first?
   ❑ always
   ❑ never
   ❑ sometimes

7. When you are writing an essay using the **computer,** do you do preplanning or a rough draft on the **computer** first?
   - ❑ always
   - ❑ never
   - ❑ sometimes

8. When you are writing an essay using the **computer,** do you do preplanning or a rough draft on **paper** first?
   - ❑ always
   - ❑ never
   - ❑ sometimes

9. Do you generally prefer writing an essay on paper or computer?
   - ❑ paper
   - ❑ computer

10. For this research, do you think the quality of the essay that you wrote on paper was different from the quality of the essay you wrote on computer?
    ❑ Y ❑ N

11. If you answered YES to question 10, please indicate which essay you felt was superior. Explain the ways in which the superior paper was better and whether you think the writing mode was a factor in the difference in quality between these two essays?

12. For this research, did your planning (prewriting, notetaking, drafting) for writing the two essays differ when you used a computer versus paper and pencil?

    ❑ Y ❑ N

    Explain your answer.

13. Did you revise your essays in the same way when using computer versus paper and pencil?

    ❑ Y ❑ N

    Explain your answer.

Appendix B

Rater Questionnaire

Rater Number _____

1. Do you own a personal computer?
   ❑ Y ❑ N

2. If yes, how long have you owned a computer?
   ❑ less than a year
   ❑ 1–3 years
   ❑ more than 3 years

3. Do you usually use a personal computer for writing papers?
   ❑ Y ❑ N

4. Do you use a computer to prepare and/or deliver instruction?
   ❑ Y ❑ N

5. How would you rate your keyboarding skills?
   ❑ poor
   ❑ just adequate
   ❑ good

6. When you are handwriting a document, do you do preplanning or a rough draft on **paper** first?
   ❑ always
   ❑ never
   ❑ sometimes

7. When you use the **computer** to write a document, do you do preplanning or a rough draft on the **computer** first?
   - ❑ always
   - ❑ never
   - ❑ sometimes

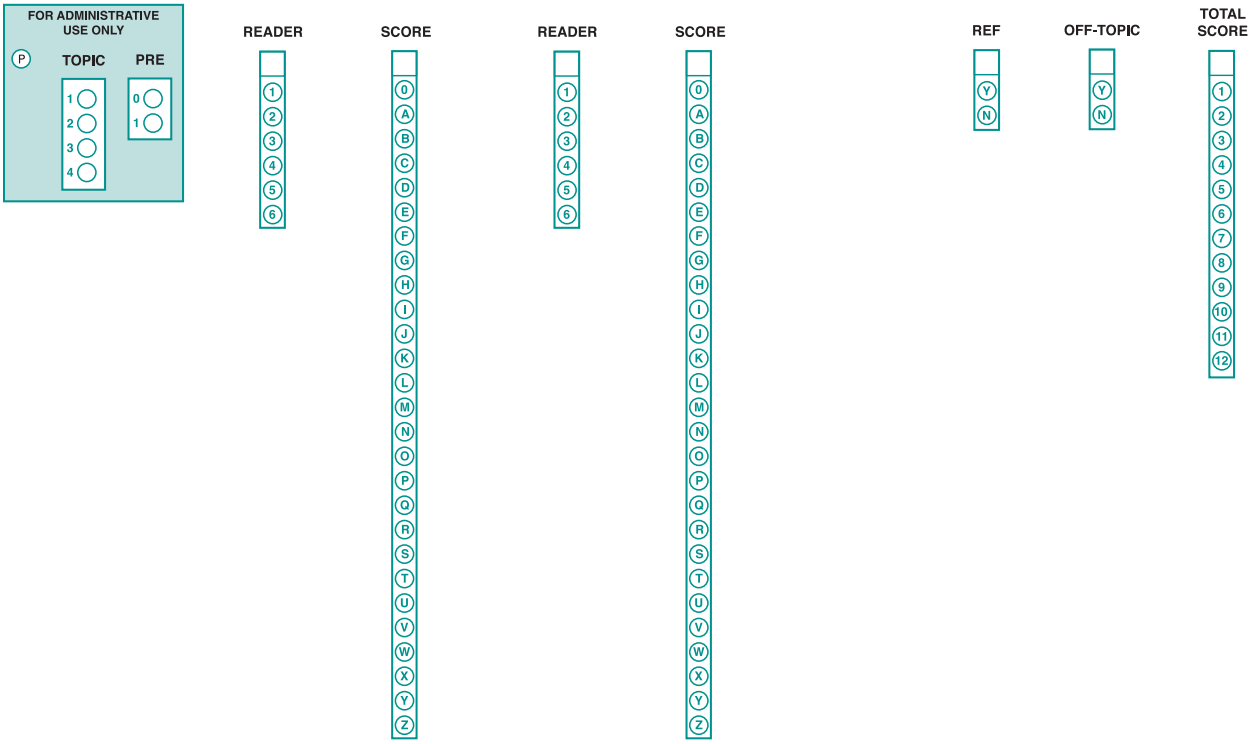
8. When you use the **computer** to write a document, do you do preplanning or a rough draft on **paper** first?
   - ❑ always
   - ❑ never
   - ❑ sometimes

9. Do you generally prefer writing documents on paper or computer?
   - ❑ paper
   - ❑ computer

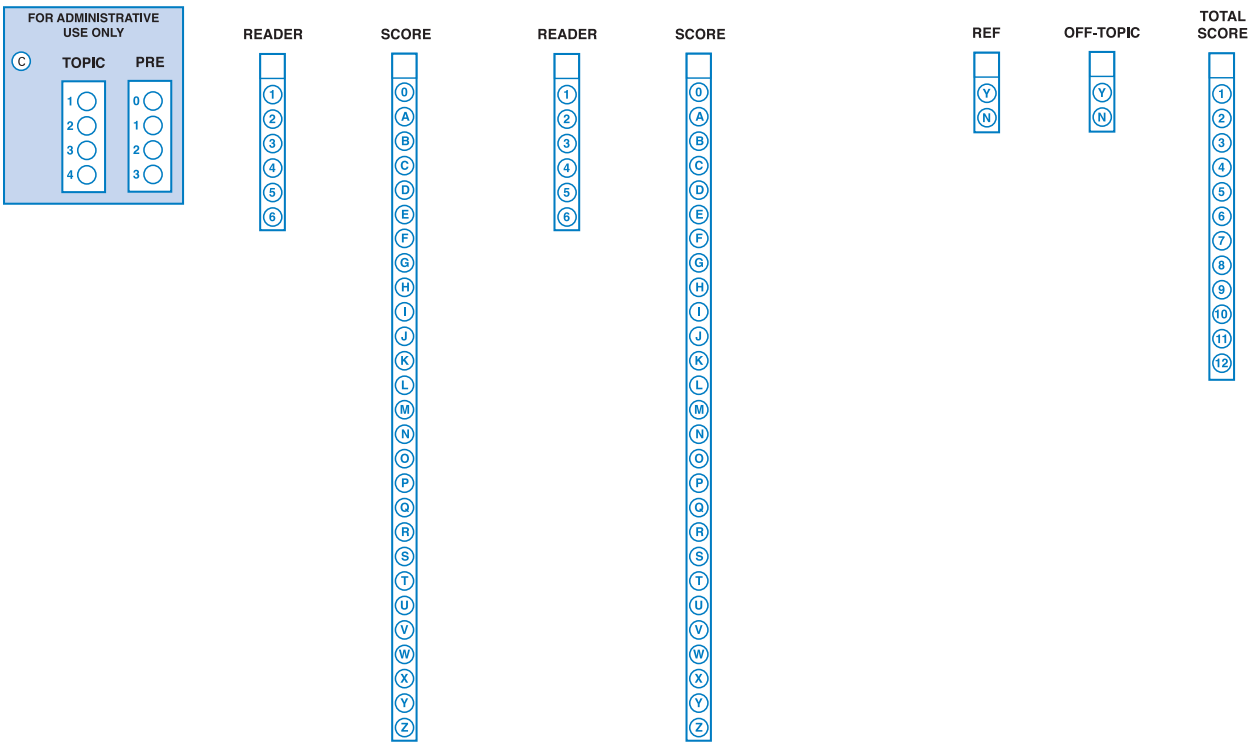
10. Do you think the quality of documents you write by hand and the quality of documents you write on computer are different?
    - ❑ Y ❑ N

11. If you answered YES to question 10, please indicate which mode of writing is generally most effective for you. Explain how you think that mode of writing affects the quality of documents you create.

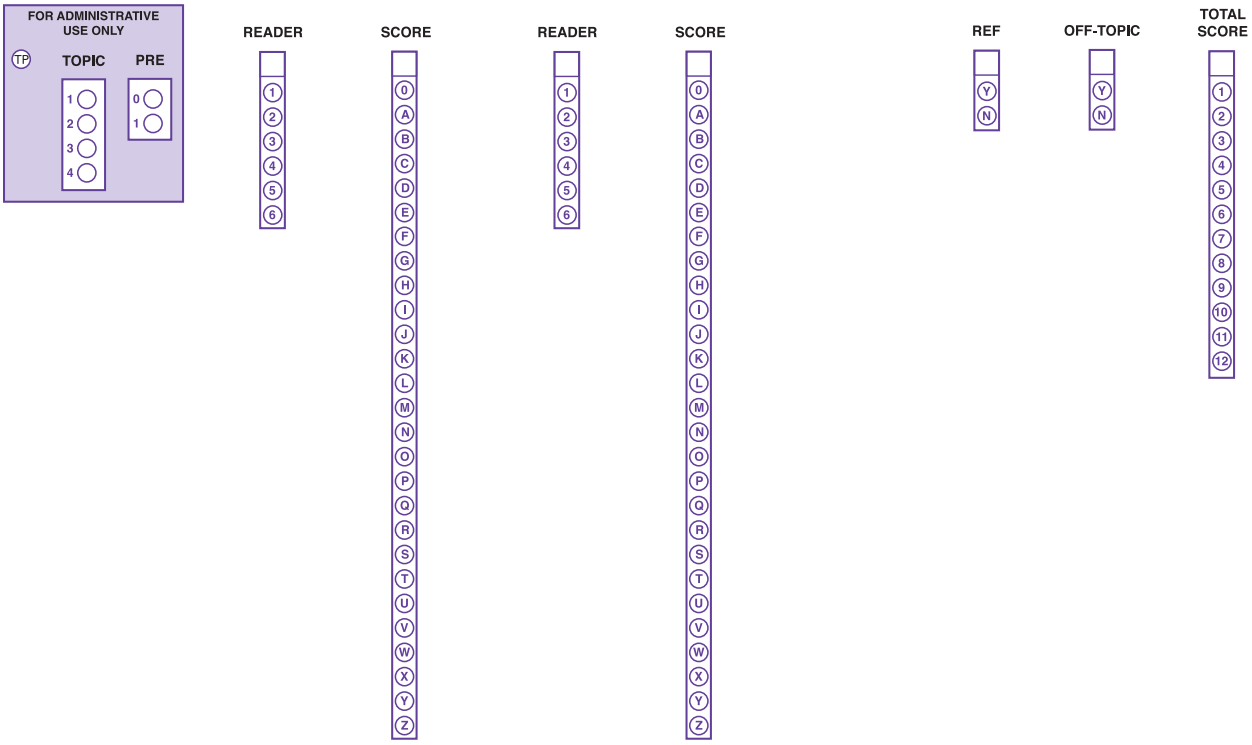
Appendix C



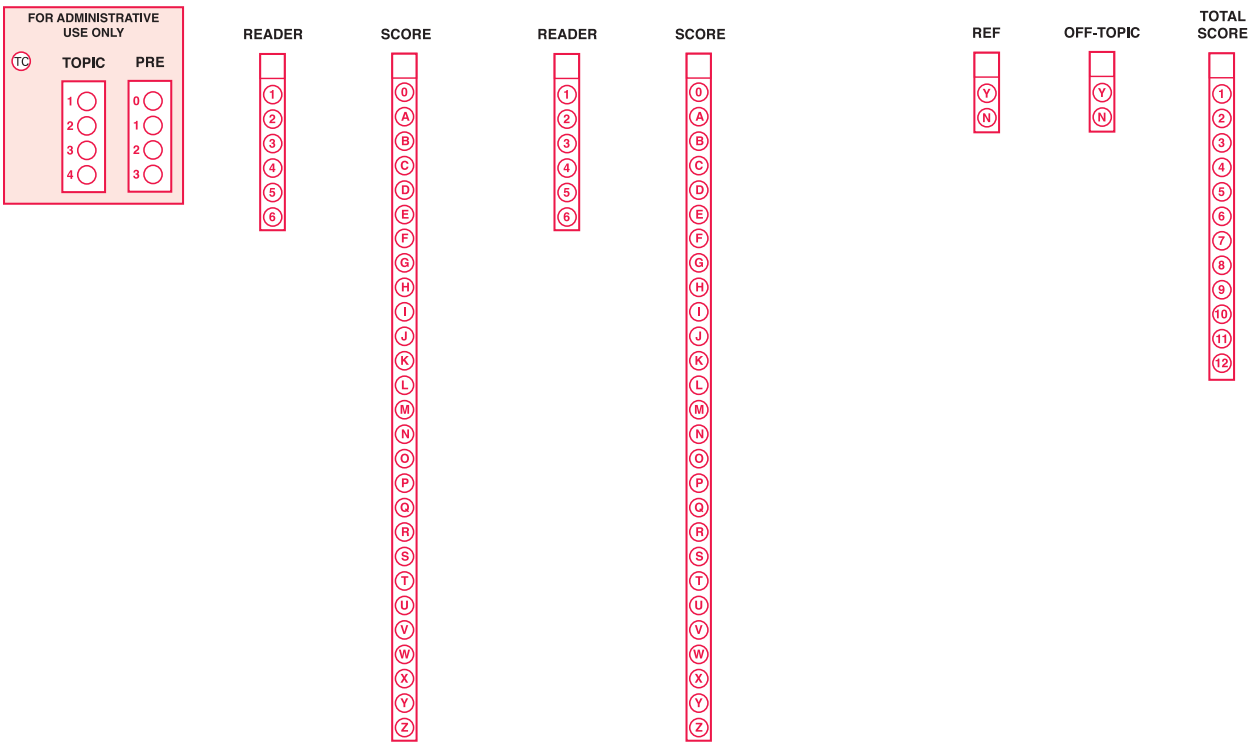
EXAMINEE ID#: _____


FOR ADMINISTRATIVE USE ONLY
C
TOPIC
1 2 3 4
PRE
0 1 2 3
READER
1 2 3 4 5 6
SCORE
0 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
READER
1 2 3 4 5 6
SCORE
0 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
REF
Y N
OFF-TOPIC
Y N
TOTAL SCORE
1 2 3 4 5 6 7 8 9 10 11 12

EXAMINEE ID#: _____


FOR ADMINISTRATIVE USE ONLY
TP
TOPIC
1
2
3
4
PRE
0
1
READER
1
2
3
4
5
6
SCORE
0 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
READER
1
2
3
4
5
6
SCORE
0 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
REF
Y
N
OFF-TOPIC
Y
N
TOTAL SCORE
1
2
3
4
5
6
7
8
9
10
11
12

EXAMINEE ID#: _____


FOR ADMINISTRATIVE USE ONLY
TC
TOPIC
1
2
3
4
PRE
0
1
2
3
READER
1
2
3
4
5
6
SCORE
0
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
READER
1
2
3
4
5
6
SCORE
0
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
REF
Y
N
OFF-TOPIC
Y
N
TOTAL SCORE
1
2
3
4
5
6
7
8
9
10
11
12

EXAMINEE ID#: _____