Available online at www.sciencedirect.com
SCIENCE @ DIRECT®
Assessing Writing 9 (2004) 4–26
ASSESSING WRITING

A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test

H.K. Lee

University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Abstract

This study aimed to comprehensively investigate the impact of a word-processor on an ESL writing assessment, covering comparison of inter-rater reliability, the quality of written products, the writing process across different testing occasions using different writing media, and students' perception of a computer-delivered test. Writing samples of 42 international students taking two tests, a paper-and-pencil based ESL placement writing test and a computer-based one, were analyzed. The results showed that while there was no significant mean difference in the holistic ratings across test occasions, all the analytic components of the computer-generated essays were marked significantly higher than those of the paper-based essays. In a holistic measurement, rater' reliability was significantly higher in word-processed essays than in paper-written essays. A student survey revealed that habitual computer writers preferred a computer-delivered writing mode to a traditional testing mode. Suggestions and implications for further study for implementing computer-based ESL writing tests are discussed.

Keywords: Assessment; English as a second language; Computer-based test

1. Introduction

Even with constant technological improvements of writing tools for teaching and testing writing skills, most university-wide placement writing tests still adhere to the traditional writing medium to collect their subjects' writing samples. In those testing contexts, the advantages of writing processors are neglected in fear

E-mail address: heelee1@uiuc.edu (H.K. Lee).

1075-2935/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.asw.2004.01.001

of potential logistical problems such as limited access to computer labs on campus and the lack of sufficient accommodating seats for many examinees on a single testing occasion. However, as computers become an authentic mode for writing, assessment of computer-delivered essays could more accurately measure actual writing skill than the paper-and-pencil based test. Thus, the delay in shifting to computers for the ESL placement writing assessment could cause concern about test validity due to lack of task authenticity and fairness for habitual computer writers. This concern would be relieved only with empirical evidence that habitual computer writers perform similarly on the computer-delivered writing test and the traditional paper-and-pencil based test.

1.1. Test validity and reliability

The concept of authenticity has been central to the study of language testing in the past several decades (Bachman & Palmer, 1996; Douglas, 2000; Lewkowicz, 2000; Messick, 1994; O'Malley & Pierce, 1995; Widdowson, 1979). It has, however, been mostly associated with the nature of test tasks. Messick (1994) advocates the view originally proposed by Arter and Spandel (1992) that authentic assessment purports to describe the processes and strategies test takers use to perform their task so as to manage the work demands properly. Accordingly, "test administrators must provide realistic contexts for the production of student work by having the tasks and processes, as well as the time and resources, parallel those in the real world" (p. 18). O'Malley and Pierce (1995) also stress that authenticity in various subject assessments must be taken into account at each stage of test administration, from constructing and implementing a test, to the grading and reporting procedures.

Side by side with the concept of authenticity lies the concept of fairness. According to the APA/AERA/NCME *Standards* (1999, p. 74), "fairness requires that all examinees be given a comparable opportunity to demonstrate their standing on the construct(s) the test is intended to measure. Just treatment also includes such factors as appropriate testing conditions and equal opportunity to become familiar with the test format, practice materials, and so forth." Disallowing computers in a writing test may deprive habitual computer writers of an equal chance to perform as well as they do in their ordinary writing circumstance. Indeed, regardless of the inconsistency of research findings about the writer's performance difference between two different writing media, habitual computer writers expressed frustration and embarrassment when given a paper-and-pencil test, as described by Collier and Werier (1995).

1.2. Studies on the effect of computers on writers

The act of writing has been addressed from various perspectives. Most importantly, it involves complex cognitive processes comprising of versatile recursive stages such as planning, translating or formulating, and reviewing (Hayes &

Flower, 1981; Kellogg, 1996; Zimmermann, 2000). In addition, writing represents a social behavior, an interplay between the creator of the message and the receiver. Good writers are capable of envisioning their audience and thereby carrying out effective communication with the audience through the text (Kroll, 1985; Rubin, 1984).

With the breakthrough of technology for writing activities, researchers have stressed the significance of technology and the physical conditions circumscribing writing tasks to the intrinsic cognitive and social perspectives of the writer (e.g., Dorner, 1992; Norman, 1989; Sharples, 1994; Warschauer, 1996). According to Daiute (1985), writing on a computer fits the well-supported notion of process-oriented writing because easier access to text enables writers to proceed continuously toward the final draft. Further, easier application of the strategy of collaborative peer editing using computers reinforces the writer's skill to perceive and correspond to social demands, leading to an increase in students' motivation for writing (see Sharples, 1994; Warschauer, 1996). In this view, therefore, writing does not represent a simple task explicable from a single point of view, but rather a complex interplay among various aspects of behaviors in which human beings engage. Among these aspects, the physical conditions surrounding writers present an important factor contributing to variance in their writing performance.

Drawing on this theoretical support for computer use in writing, a number of studies have attempted to investigate the effect of computers on writers. To date, however, no consensus has emerged. For example, in the L2 literature, word processing had no effect in some studies (Benesch, 1987; Chadwick & Bruce, 1989), while it had a positive effect on content length and quality in others (Lam & Pennington, 1995; McGarrell, 1993 cited in Pennington, 2003). This inconsistency is repeated for all population groups (professional and inexperienced writers), for writing subprocesses such as planning and revision, as well as for writing components (organization, focus, and cohesion). Collier and Werier (1995) found that professional computer writers composed similarly in paper-and-pencil based writing despite their discomfort with paper writing. In Wolfe, Bolton, Feltovich, and Niday's study (1996), students having high to medium computer experience were not affected by the writing medium, while those with little experience were adversely affected by it. In comparison of writing processes between writing on a computer and writing on paper, a number of L2 studies observed extension of the planning stage (e.g., Akyel & Kamisli, 1999; Li & Cumming, 2001), while others reported shorter planning time in computer-writing (Haas, 1989). Whereas some L2 research findings state that the revision process was enriched and extended in computer-writing (e.g., Chadwick & Bruce, 1989; Grejda & Hannafin, 1992; Phinney & Khouri, 1993), others found that writers' attention to local appearance while writing on computer impeded substantial revision (e.g., Bridwell-Bowles, Johnson, & Brehe, 1987; Collier, 1983; Haas, 1989). For the writing product quality, Pennington (1996) and Schwatz, Fitzpatrik, and Huot (1994) supported improvement in overall writing quality in computer-writing, whereas Burley (1994) observed less focus and coherence in computer-composed essays. In addition, a number of studies failed to find a consistent difference in performance between handwritten and computer-generated tests (e.g., Daiute, 1985; Hawisher, 1987; Rhodes & Ives, 1991).

Among numerous studies, Harrington's (2000) study deserves a detailed description due to its contextual similarity to the current study. In her study, 480 ESL students taking an ESL placement test were randomly assigned to one of three groups: handwriting, computer-writing and later transcribed handwriting. A mean comparison of the three groups of essays detected no significant difference, implying no disadvantage for taking computer-delivered tests. However, in addition to the group comparison, performance differences within individuals would have offered practical insights for placement judgment.

Overall, given the conflicting findings shown in various studies, particular research outcomes seem to be dependent on research contexts and specific details of the research procedures. Hence, the issue of computer-related effects on writing has to be clarified within each research context.

1.3. Raters' reactions to the word-processed text

Increasing rater reliability has been an important, but unresolved issue in writing assessment (Huot, 1990; Lumley & McNamara, 1995; McNamara, 1996). Still, in the modern era when interpretative approaches pervade, Moss (1994) argues that an individual rater's subjective context-bound judgment should not be suppressed for the sake of agreement with other raters. From any perspective, the often-reported handwriting effect on raters leading to lower reliability would not be welcomed.

Many studies have consistently addressed the association of poor handwriting with lower marks from raters (e.g., Chase, 1986; Markham, 1976). Earlier studies described that raters are likely to assign a lower grade to typed essays than to their handwritten counterparts (e.g., Arnold et al., 1990; Bridgeman & Cooper, 1988; Sweedler-Brown, 1991). Even with rater training focused on handwriting impact, raters still showed this same tendency when grading handwritten and typed essays, as shown in Powers, Fowles, Farnum, and Ramsey's study (1994). The authors attributed this tendency to a seemingly lengthier image in handwritten papers, greater conspicuity of errors in the typed essays, and raters' increased expectation in word-processing tasks. The research thus consistently supports the existence of a handwriting effect on raters. Investigating how handwriting affects readers and to what extent it does so, would be of importance not only for the current research inquiry itself, but also for the validity of the current research findings.

2. The present study

The controversial array of research findings regarding the effect of writing medium substantiates a more contextualized study. Hence, this study proposes to investigate the effect of computers on large-scale, time-constrained writing assessment at a local university. The current study comprehensively deals with issues related to computer-writing assessment including the comparison of habitual computer writers' performance in two different writing media, the investigation of the raters' reactions to different textual images, and writers' perceptions about the computer-delivered essay test. The present study is guided by several research questions.

1. How would the placement results differ between a paper-and-pencil based essay test and a computer-delivered test?
2. With analytic measurement, to what extent do feature scores vary among the different analytic features across three types of essays, handwritten, transcribed from the handwritten, and computer-delivered essays?
3. What differences are observed in the raters' rating behavior when they grade word-processed and handwritten essays?
4. How do subjects who are habitual computer writers perceive a computer-delivered essay test?

3. Method

3.1. Subjects

The subjects were solicited from the three consecutive regular paper-and-pencil based ESL Placement Tests (EPT) at the University of Illinois at Urbana-Champaign in Fall 2001. The EPT is administered to place incoming international students into an appropriate level of ESL courses and employs the traditional paper-and-pencil based testing mode. Upon the completion of each regular EPT, I advertised a chance to take the computer-delivered EPT. The incentive was offered that the better grade between the regular and the computer EPT would be considered for the final placement decision. Among 75 sign-ups, a total of 42 subjects, 16 females and 26 males participated in the computer-delivered test. Five were undergraduate students and 37 were graduate students. The subjects came from nine different native language backgrounds and from various departments.

3.2. Testing procedure

The subjects in the study took the regular EPT essay test. In the regular, paper-based EPT essay test, they wrote an essay after listening to a 10-minute lecture and reading a topic-relevant article. Fifty minutes were given for reading the article and writing their essay. The computer-delivered EPT was administered within a maximum of three days from the regular paper-based EPT in a computer lab with 28 Macintosh and 24 IBM stations. The topic for the computer EPT was 'Brain Specialization,' and the topics for the paper EPTs were 'Ethics' and 'Trade.' Specific disciplines from which the three topics were generated are economics for the topic, ‘Trade,’ neurolinguistics for ‘Brain Specialization,’ and philosophy for ‘Ethics.’ The writing prompts for the three topics (videotaped lectures and the related reading articles) currently used in the EPT essay test have been empirically validated as parallel over various subgroups of examinees despite the origin of distinctly different disciplines of the topics (Liu, 1997).

For the computer-delivered essay test, I was able to use the EPT essay testing software, previously constructed by another researcher (Kim, 2002). In the test, the subjects first faced test instructions on the computer, which had been transcribed from the cover page of the paper test booklet. They then watched the videotaped lecture on the computer with headphones on. The listening stimulus was computer-delivered for equal aural effect to all examinees. As in the regular EPT, the video was shown only once and note-taking was allowed during listening. After listening to the lecture, subjects were given 50 minutes to read the hard-copy article and to write their essay on the computer. In their writing, subjects could use word processing functions such as copy and paste, delete, undo, and indentation. The final essays were submitted to the server. Subsequent to the test, subjects were asked to complete a questionnaire (Appendix A) on their writing habits and processes.

As described, I designed the procedure for the computer-delivered EPT as similarly as possible to the regular paper-and-pencil based EPT, except for the difference in the writing medium and the individually delivered video prompt.

3.3. Scoring procedure

The subjects’ essays produced in the paper and computer EPT were first graded holistically at the operational scoring session held on the day of the test, and later analytically graded for this research purpose. The operational EPT scoring procedure is as follows: Each essay is independently rated by two raters based on holistic grading benchmarks. All raters are ESL teaching assistants with more than a semester of teaching experience, and all participate in a mandatory training session. The essays are marked with one of the four placement levels: too low, ESL 400, ESL 401, and exempt for graduate level and, too low, ESL 113L, ESL 113U, and ESL 114 for undergraduate level.[1] In case of a one-level score difference between the two ratings for a particular essay, raters discuss the essay to reach a consensus. A score discrepancy of two levels is resolved by a third rater. Appendix B gives the holistic benchmarks used in the operational EPT grading.

Subsequently after the computer-delivered test, I retrieved the 42 computer-written essays from the server and printed for scoring. In advance to the computer-delivered testing date, I had word-processed the 42 paper-written essays produced

[1] The score, ‘too low,’ is rarely assigned to essays. There is no ‘exempt’ score for undergraduate students because the university strictly regulates rhetoric requirement for all undergraduate students. For international undergraduate students, taking sequential classes of ESL 114 and 115 is considered to fulfill the rhetoric requirement.

by the 42 subjects at the paper-based test (this version is subsequently called ‘transcribed’), and mixed them with the 42 computer-written essays for grading. The 42 transcribed essays had been typed literally with any spelling or grammatical errors left intact and the textual quality and representation format preserved. Essays were identified by a series of numbers after removal of writers’ names. I did not inform raters that the essays included the transcribed set, so that raters could regard all typed essays as produced in the computer EPT. The computer-written and transcribed essays were operationally graded on the day of the computer-delivered essay test. A team of eight qualified raters participated in the operational scoring session.

Later, the three sets of 126 essays (handwritten, transcribed, and computer-written) were rated analytically based on the feature analysis form (Appendix C). I developed this form with close reference to the operational EPT holistic benchmarks: For the analytic rating, I used the same four placement levels to make them parallel to the ones in the holistic benchmarks. I extracted four features from the holistic benchmarks, and copied the descriptors of each level of the four features directly from each corresponding level of the holistic benchmarks.

Two trained ESL teachers with more than one year of teaching experience worked on the analytic rating of all the essays. Again, I did not inform them of the existence of the transcribed version. To avoid a recall effect for essay content, they were initially given the 42 handwritten essays, then the 42 computer-written essays. Finally, they rated the 42 transcribed essays after the first and second sets of essays were collected. The period of analytic rating spanned a full month, so the chance of remembering the content of a particular handwritten essay was minimized.

4. Results and discussion

4.1. Inter-rater reliability

The percentage of exact agreement in the holistic measurement of the handwritten essays was 64.3%, that is, 27 out of 42 handwritten essays were assigned the same placement level by two raters. On the other hand, 33 (78.6%) out of the 42 transcribed essays and 32 (76.1%) out of the 42 computer-written essays produced exact agreement between two raters. All the other essays were disagreed by two raters just by one placement level. Inter-rater reliability was computed between two original scores given by two independent raters per essay. Table 1 compares the inter-rater reliability among the three types of essays for two assessment occasions, holistic, and analytic assessment. For an inter-rater reliability index, the Intraclass Correlation Coefficients (ICC) were calculated using 11.0 Version of SPSS.

A test of the difference in the correlations between the holistic scores of handwritten and transcribed essays yielded statistical significance at $z = 2.70$, $P <$ .005. The same held true in comparison of the correlations between the handwritten

Table 1
Inter-rater reliability

Type of essays	Reliability in the holistic measurement	Reliability in the feature analysis
Handwritten essays	.60	.79
Transcribed essays	.86	.70
Computer-generated essays	.81	.68

and computer-generated essays at $z = 1.96$, $P < .05$. These results suggest that word-processed essays are more resistant to discrepancies in score judgment between readers. Further, considering the significantly lower reliability of handwritten essays, and the rather low reliability index of the handwritten essay set, handwritten essays seem to have considerable handwriting impact on raters.

On the other hand, reliability in the feature analysis was consistent for all essay types, with non-significant differences in the correlations between any two essay types. This consistent rater reliability over all essay types shown in the analytic grading might stem from the fact that, unlike the holistic measurement based on an impressionistic judgment, the nature of feature analysis requires more focused and longer reading as observed in Spandel and Stiggins' study (1980). Alternatively, this tendency could be due to the small number of raters, which was only two, involved in the feature analysis process, compared to eight readers participating in holistic assessment. The fact that the reliability figures are moderated across the essay type ensures the dependability of feature analysis rating and greater impartiality of the raters with respect to handwriting.

In summary, the significant difference in inter-rater reliabilities between handwritten and word-processed texts in the holistic assessment suggests that in the operational scaling, handwritten texts do not yield consistent decisions from readers. In accordance with the current notion of reliability as subsumed into a validity argument (Chapelle, 1999), and the notion that individual raters are an important facet in writing assessment, a test resulting in less agreement between from raters is difficult to support.

4.2. Holistic measurement

As noted, all the essays in this study were given one of four placement levels. For statistical purposes, the scores from ‘too low’ to ‘exempt’ were converted to a numerical scale of one to four, respectively. Table 2 presents the descriptive statistics of the three types of essay scores in the holistic measurement.

The repeated-measures ANOVA detected no significant mean difference in the holistic scores of the three types of essays. Since in the placement tests a more meaningful interpretation can be drawn from the discrepant cases of placement results within individuals on different occasions, a comparison was made of placement results per subject assessed by different essay type. Table 3 compares the

Table 2
Descriptive statistics for the holistic measurement ($n = 42$)

Essay type	Mean	*SD*	Minimum	Maximum
Paper	2.55	0.50	2.00	3.00
Transcribed	2.60	0.66	2.00	4.00
Computer	2.71	0.71	2.00	4.00

placement results in the computer-delivered test with those in the paper-and-pencil based test.

As indicated, there was a considerable discrepancy regarding the placement results based on the computer-written essays and handwritten essays. Specifically, among 42 subjects, seven (16.7%) received an improved score by one level, and two (4.8%) were placed two levels above their original placement by the paper-based EPT. On the other hand, four subjects (9.5%) received lower placement results by one level. On the whole, the placement results were enhanced by the computer-based EPT. Surprisingly, two subjects were upgraded by two levels by the computer EPT. For those two subjects, the writing tools affected their performance to a great degree, and thus it seems hard to argue that only the testing mode made these differences. Admittedly, the practice effect seemed to come into play in the second test, as shown in subjects' survey comments. The two subjects confessed that in the second test they could take more control of the given time than in the first test. Nevertheless, surveys revealed that subjects still favored the computer testing conditions, and believed them to be the main factor in their improved performance.

On the other hand, four subjects got a lower grade on the second test, and the cause must be investigated in depth. The analysis of survey responses by these subjects revealed that two of them took the test simply because they wanted to have a chance to compensate for their first results, disregarding their preference for the handwriting medium. In those cases, non-authentic and unfamiliar writing instruments in the test resulted in diminished performance. As Wolfe et al. (1996) found, people with little computer experience performed worse in computer-based writing than they did in the traditional writing mode. The other two subjects reported

Table 3
Frequency of placement result in the computer EPT and paper EPT

	Computer			
Paper	ESL 400	ESL 401	Exempt	Total
ESL 400	14	3	2	19
ESL 401	4	15	4	23
Exempt	0	0	0	0
Total	18	18	6	42

Note. Scores of five undergraduate students were included in the analysis by placing their score corresponding to the score of graduate level. For example, the undergraduate placement result, 'ESL 114' coincided with the corresponding graduate score, 'exempt.'

Table 4
Frequency of placement results between the handwritten essays and the transcribed essays

	Typed			
Hand	ESL 400	ESL 401	Exempt	Total
ESL 400	15	4	0	19
ESL 401	6	13	4	23
Exempt	0	0	0	0
Total	21	17	4	42

that the topic in the second test seemed more difficult than that in the first, regular EPT. As topic is widely perceived as one of the significant factors affecting writers, it appeared that a conceivably more difficult topic diminished the advantage of using familiar writing tools. In sum, although the higher mean score of the computer-delivered test did not result in a statistically significant difference from the paper-based EPT, an examination of individual cases substantiates performance differences in tests with different writing media.

Although the holistic scores on the transcribed version of essays were no longer important for the practical placement decision, from an empirical point of view a comparison of scores of the pairs of transcribed essays and original handwritten essays would shed light on the validity of the current study. As readers are one of the inevitable sources of error in writing assessments, and text appearance has been often reported to affect raters, an examination of the handwriting effect on raters is necessary to check the validity of the assigned scores.

Table 4 displays the holistic scores between pairs of original handwritten essays and their transcribed counterparts. In the measurement of the 42 transcribed essays, 8 essays were scored higher by one level, whereas 6 were marked one level lower than the handwritten counterparts. Thus, the mean difference between the transcribed and handwritten versions was slight, with the transcribed essays being marked higher. This result contradicts outcomes from earlier studies such as Arnold et al. (1990) and Powers et al. (1994) in which transcribed essays got higher marks than handwritten essays.

The contrary finding in the current study seems likely due to the severe time-constraints imposed on raters. As scores must be available by the next business day after testing, raters need to finish their work between 11.00 a.m. and 5.00 p.m. of the test day. Furthermore, there is a great influx of the EPT test takers each fall semester, and the limited number of raters must handle 80–100 essays in this limited time. Under such time pressure, as Sloan and McGinnis (1978) also reported, readers evaluating many essays as rapidly as possible tended to assign lower scores to messy handwritten essays than to neat ones because raters who had to pass essays quickly to another reader were not able to devote a sufficient amount of time to deciphering messy handwriting.

Indeed, our raters agreed in the follow-up interview that when they read severely illegible essays, they had been more likely to give low scores to the essays. The raters added that the illegible essays presented 15–20% of the total number of essays, and they interrupted the smooth flow of reading and impaired their focus on content. Therefore, in a time-constrained testing condition for writers and raters alike, handwriting may play a negative role.

Regardless of the non-significant statistical outcome, the placement results in the handwritten essays and transcribed counterparts in the discrepant 14 cases would warrant substantial consideration. About a third of subjects who might have been placed at a different level had their essays been written on a computer would get disservice from their writing class simply due to raters' error. Extending these results to the entire EPT population, this is a harmful factor impairing the validity of the EPT. Furthermore, this result prompts a reconsideration of the comparison of results between computer essays and handwritten essays analyzed above. Admittedly, the existence of the handwriting effect shown in the 14 cases suggests that the better scores on the computer-written essays might have been confounded with the handwriting effect along with subjects' real performance differences.

To conclude, for a third of all the essays, the measurement of the transcribed essays differed from that of the handwritten counterparts. This finding calls our attention to rater training. It strongly suggests that test administrators need to be alert and train raters to focus on just the content and immunize themselves to text appearance. Rater training will be re-addressed in a later discussion.

4.3. Analytic assessment

Descriptive statistics for the feature analysis are presented in Table 5. The feature scores of each essay were the average of the marks assigned by two raters. Table 5 shows that across all features, essays in the computer-based EPT displayed the highest means, and the essay pairs of paper and transcribed versions had similar means.

The repeated-measures MANOVA revealed highly significant test type effects ($P < .001$) for the features (Wilks' Lambda $= 0.504$, $F = 4.187$ with $df = 8$). In repeated-measures MANOVA, it is important to satisfy the assumption of sphericity, equality of variances of differences for all pairs of levels of the repeated-measures factor. This is tested by Mauchley's sphericity test. The absence

Table 5
Descriptive statistics in the feature analysis ($n = 42$)

	Paper EPT		Transcribed		Computer EPT	
Feature	*M*	*SD*	*M*	*SD*	*M*	*SD*
Organization	2.35	0.46	2.32	0.52	2.65	0.45
Content	2.70	0.46	2.69	0.48	2.98	0.53
Use of sources	2.54	0.57	2.49	0.57	2.81	0.60
Linguistic expression	2.48	0.52	2.39	0.48	2.63	0.56

Table 6
Univariate ANOVA for each feature

Source	Measure	SS	*df*	MS	*F*	*P* > *F*
Essay	Organization	2.90	1.71	1.70	11.42	.000
	Content	2.20	1.75	1.25	9.18	.001
	Source use	2.53	1.49	1.70	5.93	.009
	Linguistic expression	1.23	1.62	1.62	5.33	.011

of sphericity can be adjusted by lowering degrees of freedom. Greenhouse–Geisser test produces corrected degrees of freedom for the adjustment. In the current data, as Mauchley's test of sphericity detected significance at $\alpha = 0.05$ across the four features, the degrees of freedom were adjusted with the values of the Greenhouse–Geisser test in the follow-up univariate ANOVA. Table 6 reports the univariate tests for each feature.

As all features showed a significant essay type effect, post hoc pairwise contrasts were conducted. Table 7 summarizes tests of contrasts. As seen, the comparison of each feature between the transcribed and the handwritten groups yielded no statistically significant mean difference, which is desirable in writing measurement. In other words, regardless of different textual image, raters made judgments on the quality of feature components in a consistent way. Early in this section, the raters' reliability was also exhibited by the moderate degree of inter-rater reliability in the feature analysis corresponding to all three essay types.

On the other hand, the tests of contrasts detected significance in the mean difference between all features of the handwritten and computer-delivered essays. All features of essays produced on the computer EPT were scored significantly better than those of the essays produced on the paper EPT. To reiterate, as far

Table 7
Tests of pairwise contrasts for the four features

Feature	Essay	Mean difference	*P* > *F*
Organization	COM versus TRAN	0.33	.001
	COM versus PAP	0.31	.000
	PAP versus TRAN	0.02	.700
Content	COM versus TRAN	0.29	.000
	COM versus PAP	0.27	.003
	PAP versus TRAN	0.01	.850
Use of sources	COM versus TRAN	0.32	.003
	COM versus PAP	0.27	.032
	PAP versus TRAN	0.05	.499
Linguistic expression	COM versus TRAN	0.24	.008
	COM versus PAP	0.15	.050
	PAP versus TRAN	0.08	.128

Note. COM: computer-delivered essays; TRAN: transcribed essays; PAP: paper-and-pencil based essays.

as the feature scores were concerned, subjects performed better overall on the computer-delivered essay test than on the paper-based test. The degree of significance of the mean difference was see to be highest in the organization feature, followed by content.

Predictably, organization was enhanced in the computer-based domain, likely due to the extended time for planning and review as in Kellogg's study (1994). This claim was supported by subjects' responses to the survey item 21, which asked if they could organize better on the computer EPT. Approximately 60% of subjects marked on the positive two scales for this item.[2] Additionally, half of the subjects thought that reviewing the text was easier than in the paper EPT.

Contrary to the findings of Freedman and Clarke's study (1988) in which computer writers made no improvement in the overall content due to the greater efforts they devoted to correcting local and surface errors, this study discovered enhanced content with computer-writing. The improved content might be ascribed to reasons similar to those influencing the organization feature. Better structuring and reviewing of the text appeared to result in better content. It is also possible that, as Rodrigues (1985) demonstrated, the highly readable screen display might encourage more reading of the text, resulting in more in-depth revision. Further, over half the subjects said that they could write more sentences and thus expected better results. Although quantitative text-increase did not necessarily indicate a qualitative improvement, adding text may tend to enrich the content overall.

The enhancement of the linguistic expression feature with computer-writing seems to imply that subjects did not pay attention merely to the local level of errors. Sixty percent of the subjects answered positively that in the computer EPT their writing was enhanced by the opportunity for correction and revision. Further, owing to a more simplified revision process than on paper writing, subjects were able improve their writing sample by replacing awkward expressions during the revising process.

On the other hand, it is very hard to explain the more effective use of sources in the computer-written essay. The computer-writing process, facilitated throughout the task by word-processing tools, might contribute to making more use of sources. Perhaps due to the alleviated physical labor involved in writing and revising by computer, the subjects were able to read and re-read the given article. In addition, the individually delivered video prompt could give rise to a more focused listening activity, better comprehension, and therefore more facility with using the sources.

Finally, some discussion is required to explain the fact that while the holistic grading did not yield significant mean differences between computer-delivered and paper-based essays, analytic rating resulted in significant improvement of all features with the computer-delivered EPT. It is likely that impressionistic holistic ratings measured different things from analytic assessment. Although the feature analysis forms were developed based on a close consideration of holistic bench-

[2] Subsequently, when survey results are discussed, agreement rate is measured on marks on two positive scales and disagreement rate on two negative scales.

marks, raters' judgments in the holistic scale may encompass something not scrutinized in single features. Or, as Hayes, Hatch, and Silk (2000), Huot (1990), and Vaughan (1993) cast doubt on the reliability of holistic measurement, differences might be ascribable to greater rater error involved in holistic assessment.

Given the moderate inter-rater reliability for all essay types with feature analysis, and the nature of feature analysis involving a more thorough and focused reading, the results of the feature analysis deserve more emphasis in evaluating writers' performance on both the computer-delivered and paper-and-pencil based essay tests. Taken together, this study outcome suggests that subjects perform better on the computer EPT than they do on the paper-based EPT.

4.4. *Student survey*

Upon completion of the computer-delivered test, subjects responded to a five-point Likert-scale questionnaire and two open-ended questions (Appendix A). Regardless of their score on the computer-delivered EPT, the survey results revealed that subjects preferred the computer-delivered EPT to the paper-and-pencil test, as most subjects (37 out of 42) in the study agreed on their dependence on the computer as a writing medium. Notably, the subjects reported the cumbersome nature of the process of correcting and editing their text in the paper EPT. They complained of time consumption in the editing process in paper writing and the difficulty of focusing on content development. They believed that they performed better when using the computer, and expected higher scores.

In response to questions on their perception of the difference in the writing subprocesses between paper and computer-writing, more than half of the subjects felt the processes differed. However, according to subsequent responses to specific questions about where the process differences occurred (questions 12–15), no consistent pattern was found. How two writing processes differed seemed to be an individual matter.

Subjects made comments on various issues in answer to the open-ended questions. Time-constraints emerged as the main problem. Most subjects complained of the shortage of time. Some confessed that they were better able to control their time on the computer-based test, even though the same amount of time was given as in the first test. However, the source of their perception about less time pressure in the second test was unclear. There may have been a practice effect, or the use of the more familiar writing medium, a computer, may have influenced their perception. In addition, the topic of the EPT was another complaint. Subjects stated that the topics were too specific to a certain discipline to ensure fairness across all examinees. This perception of the topic was also displayed in the responses to relevant items on the survey. For survey item #11, 62% of the subjects felt that the difficulty of the topic was not the same in the two tests. Even though the three topics currently used in the EPT have been argued as empirically parallel (Liu, 1997), the subjects' perception of the level of difficulty seemed to vary across topics.

In addition, the subjects made good comments on technological matters. They preferred simpler software in the essay test. Some subjects wanted access to more advanced technologies such as spell-check and an on-line dictionary. In light of the principle of the EPT essay test, i.e., to measure an authentic, academic writing task, provision of those functions should be taken into consideration in implementing the computer-delivered test to better simulate the actual writing situation.

Overall, the survey results supported the rationale of the computer-delivered essay test: increased authenticity of the writing environment and the chance for a better performance. Consequently, subjects believed the computer test to be more valid in placing them into proper ESL classes (questions 16 and 17). This seems to indicate that the computer-delivered essay test is at least face-valid for the subjects.

5. Limitations of the study

Several limitations of the study should be noted. First of all, admittedly, the better performance on the computer EPT might have been conflated by the order effects present in the study. Indeed, some subjects commented on the survey that the time was more manageable in the second test. Also, the second test could not avoid the subjects' increased awareness of the test format and practice. Regrettably, because of logistical problems involved in the subject solicitation procedure, I could not counter-balance the order of test occasions in the current study.

Second, although I tried to minimize the effect of topics, the subjects' different perceptions of topic difficulty in the two tests might have confounded the findings. This point was also supported by the subjects' responses to the survey. Still, it is not clear, and is open to further study, whether or not the level of difficulty perceived by writers indeed affects test performance, and to what extent topics affect writers' performance.

Finally, because the subjects were all volunteers, and their number was small, generalization of the study findings needs special care. When soliciting subjects, I placed emphasis on their familiarity with computer-writing. Hence, the subject group was not representative of EPT test takers as a whole, but probably of habitual computer writers. Nevertheless, as revealed in the survey, it could not be assumed that the subjects were randomly selected from a population of computer writers, because some of them admitted that they were not, but rather took the computer test for other reasons. Therefore, the study results would not be applicable to all computer writer groups or to the general population of EPT task takers.

6. Further implications

This study brings rater training to particular attention. To date, the rater recalibrating procedure for most placement tests has not included a lesson for the possible impact of textual image. Given the significantly lower figure of inter-reliability in rating handwritten essays than in scoring word-processed essays, practical tips should be explored. For example, during recalibration, raters could be given pairs of messy handwritten and typed essays with a certain time interval to compare the scores they assign to the pair. Plenty of practice and self-calibration would immunize raters against the handwriting effect. More specific rating guidelines should be further devised for each testing context.

As shown by subjects' suggestions in the survey, the basic concerns requiring prompt considerations by test administrators were to allow more time in the test and to develop fair topics for examinees from diverse disciplines. Even though in the current EPT the time matter was resolved by the development of a process-oriented Enhanced EPT (Cho, 2001),[3] still the Enhanced EPT places limitations on the number of students that can be accommodated in a single test operation. Since the regular EPT is the primary test instrument in UIUC ESL placement procedure, the time limit needs adjustment within the regular EPT context.

The need for more and fairer topics in the EPT context has continuously emerged both from many researchers and examinees. The three current prompts designed to simulate academic writing tasks were generated from specific fields of study. Thus, the current topics might be neither general nor overly familiar to many of the examinees. Although parallelism of three prompts was empirically assured (Liu, 1997), it is not supported by the subjects' perceptions. Suggestions made in the subjects' survey indicate that using a globally general topic or using different topics specific to the various discipline groups could be possible solutions. Yet, the kind of topic that would be desirable for writing assessment still remains at the core of debate.

Although it was non-operational and only motivated by research, the analytic measurement provided richer and more substantial information about essay quality. In the future, making parallel use of analytic and holistic assessments could enhance the raters' awareness of their rating task and even improve washback to test takers and to teachers. It would also offer future researchers precious opportunities to study students' writing tendencies and provide ESL teachers with diagnostic information about ESL writers' weaknesses and strengths for teaching guidance. Future research needs to address issues such as classifying features, defining scales and descriptors, the feasibility of the analytic scoring in the current EPT context, and the analysis of feature scores compared to a single holistic score.

In implementing the computer EPT, many issues warrant further investigation. For example, it would be important to investigate the relationship between the degree of other computer skills and writing performance in the EPT context. The impact of the orientation to a particular writing domain on writing performance is another matter of investigation. Further questions would involve: consideration of technical concerns; choice of software in the particular writing field; examinees'

[3] The enhanced EPT is designed based on the notion of process-oriented writing. Students are given approximately five and a half hours for their writing activities including brainstorming, peer-feedback, self-evaluation, and rewriting.

access to functions such as spell- and/or grammar-check, and a thesaurus. Most importantly, a decision on operationalization of the computer EPT should be made with enough consideration of the fact that not all the international students are habitual computer writers and, for those who prefer writing on paper, the traditional testing format still has to be kept.

Acknowledgments

Grateful thanks are expressed to the EPT Supervisor, Dr. Fred Davidson. I am also grateful to two anonymous Assessing Writing reviewers for their valuable comments. The remaining errors are solely my responsibility.

References

Akyel, A., & Kamisli, S. (1999). Word processing in the EFL classroom: Effects on writing strategies, attitudes, and products. In: M. C. Pennington (Ed.), *Writing in an electronic medium: Research with language learners* (pp. 27–60). Houston: Athelstan.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington DC: American Educational Research Association.

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

Arter, J., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice, 11* (1), 36–44.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Benesch, S. (1987). *Word processing in English as a second language: A case study of three non-native college students.* (Available ERIC: ED281383.)

Bridgeman, B., & Cooper, P. (1988). *Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Bridwell-Bowles, L., Johnson, P., & Brehe, S. (1987). Composing and computers: Case studies of experienced writers. In: A. Matsuhashi (Ed.), *Writing in real time: Modeling composing processes* (pp. 81–107). Norwood, NJ: Ablex.

Burley, H. (1994). *Postsecondary novice and better than novice writers: The effects of word processing and a very special computer assisted writing lab*. Paper presented at South-western Educational Research Association Meeting, San Antonio, TX.

Chadwick, S., & Bruce, N. (1989). The revision process in academic writing: From pen & paper to word processor. *Hong Kong Papers in Linguistics and Language Teaching, 12*.

Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19,* 254–272.

Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement*, *23*, 33–41.

Cho, Y. S. (2001). *Examining a process-oriented writing assessment in a large-scale ESL testing context*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Collier, R. (1983). The word processor and revision strategies. *College Composition and Communication, 34*, 149–155.

Collier, R., & Werier, C. (1995). When computer writers compose by hand. *Computers and Composition*, *12*, 47–59.

Daiute, C. (1985). *Writing* & *computers*. Addison-Wesley.

Dorner, J. (1992). Authors and information technology: New challenges in publishing. In: M. Sharples (Ed.), *Computers and writing: Issues and implementation* (pp. 5–14). Dordrecht: Kluwer Academic Publishers.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Freedman, A., & Clarke, L. (1988). *The effect of computer technology on composing processes and written products of grade 8 and grade 12 students (Education and Technology Series)*. Toronto: Ontario Department of Education.

Grejda, G. F., & Hannafin, M. J. (1992). Effects of words processing on sixth graders' holistic writing revisions. *Journal of Educational Research*, *85*, 144–149.

Haas, C. (1989). *How the writing medium shapes the writing process: Effects of word processing on planning*. (Available ERIC: EJ388 596.)

Harrington, S. (2000). The influence of word processing on English placement test. *Computers and Compositions*, *17*, 197–210.

Hawisher, G. E. (1987). The effects of word processing on the revision strategies of college freshman. *Research in the Teaching of English*, *21*, 145–159.

Hayes, J. R., & Flower, L. S. (1981). Identifying the organization of writing process. In: L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hayes, J. R., Hatch, J. A., & Silk, C. M. (2000). Does holistic assessment predict writing performance? *Written Communication*, *17*, 3–26.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*, 201–213.

Kellogg, R. T. (1994). *The psychology of writing*. New York, NY: Oxford University Press.

Kellogg, R. T. (1996). A model of working memory in writing. In: C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–72). Mahwah, NJ: Lawrence Erlbaum.

Kim, J. (2002). *Computer-delivered ESL Writing Placement Test at the University of Illinois at Urbana-Champaign*. Unpublished masters thesis, University of Illinois at Urbana-Champaign.

Kroll, B. M. (1985). Social-cognitive ability and writing performance: How are they related. *Written Communication*, *2*, 293–305.

Lam, F. S., & Pennington, M. C. (1995). The computer vs. the pen: A comparative study of word processing in a Hone Kong secondary classroom. *Computer-Assisted Language Learning*, *7*, 75–92.

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, *17*, 43–64.

Li, J., & Cumming, A. (2001). Word processing and second language writing: A longitudinal case study. In: R. Manchon (Ed.), *Writing in the L2 classroom: Issues in research and pedagogy*. *International Journal of English Studies*, *1*, 127–152.

Liu, H. (1997). *Constructing and validating parallel forms of performance-based writing prompts in an academic setting.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 54–71.

Markham, L. R. (1976). Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal*, *13*, 277–283.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher*, *23*, 13–23.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*, 5–12.

Norman, D. A. (1989). Cognitive artifacts. In: J. M. Carroll (Ed.), *Designing interaction: Psychology at the human-computer interface* (pp. 17–38). New York: Cambridge University Press.

O'Malley J. M., & Pierce, L. V. (1995). *Authentic assessment for English language learners, practical approaches for teachers*. Addison-Wesley.

Pennington, M. C. (1996). *The computer and the non-native writer: A natural partnership*. Cresskill, NJ: Hampton Press.

Pennington, M. C. (2003). The impact of the computer in second language writing. In: B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 287–310). Cambridge, UK: Cambridge University Press.

Phinney, M., & Khouri, S. (1993). Computers, revision, and ESL writers: The role of experience. *Journal of Second Language Writing, 2*, 257–277.

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

Rhodes, B. K., & Ives, N. (1991). *Computers and revision — Wishful thinking or reality*? (Available ERIC: ED 331 045.)

Rubin, D. L. (1984). Social cognition and written communication. *Written Communication, 1*, 211–245.

Schwatz, H., Fitzpatrik, C., & Huot, B. (1994). The computer medium in writing for discovery. *Computers and Composition, 11*, 137–149.

Sharples, M. (1994). Computers support for the rhythms of writing. *Computers and Composition, 11*, 217–226.

Sloan, C. A., & McGinnis, L. (1978). *The effect of handwriting on teachers' grading of high school essays*. (Available ERIC: ED 220 836.)

Spandel, V., & Stiggins, R. J. (1980). *Direct measures of writing skill: Issues and applications*. Portland, OR: Northwest Regional Educational Development Laboratory.

Sweedler-Brown, C. O. (1991). Computers and assessment: The effect of typing versus handwriting on the holistic score of essays. *Research and Teaching in Developmental Education, 8*, 5–14.

Vaughan, C. (1993). Holistic assessment: What goes on in the rater's mind? In: Hamp-Lyons, L. (Ed.), *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation.

Warschauer, M. (1996). Motivational aspects of using computers for writing and communication. In: Warschauer, M. (Ed.), *Wtelel collaboration in foreign language learning* (pp. 29–46). Honolulu, HI: University of Hawaii Second Language Teaching and Curriculum Center.

Widdowson, H. G. (1979). *Explorations in applied linguistics*. Oxford: Oxford University Press.

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

Zimmermann, R. (2000). L2 writing: Subprocesses, a model of formulating and empirical findings. *Learning and Instruction, 10*, 73–99.

Appendix A. Student survey form

Last Name: _______________ First Name: ________________ SSN: ___________________

Directions: This questionnaire is given to you to find out general writing process and test-taking experiences. Read each question carefully and choose the choice that best represents your opinion.

		Strongly disagree			Strongly agree	
1	I liked the computer EPT better than the paper-and-pencil based EPT.	1	2	3	4	5
2	I could write better in the computer EPT than at the paper EPT.	1	2	3	4	5
3	I felt more comfortable at the computer EPT.	1	2	3	4	5
4	The computer EPT provided an easier writing situation than the paper EPT.	1	2	3	4	5
5	Writing on a computer represents my habitual writing situation better than the paper EPT.	1	2	3	4	5
6	Most of the time, I type on a computer when I cope with my writing workload.	1	2	3	4	5
7	I usually use both writing methods (writing on a paper and on a computer) when I compose an essay.	1	2	3	4	5
8	I think my writing process differs when I write on a computer from when I do on a paper.	1	2	3	4	5
9	I usually write better when I write on a computer.	1	2	3	4	5
10	In my case, the writing tool does not make any difference in terms of the writing quality.	1	2	3	4	5
11	I think the difficulty of the topic was the same in the two tests.	1	2	3	4	5
12	At the computer EPT, I spent more time organizing ideas than at the paper EPT.	1	2	3	4	5
13	I spent more time translating my ideas into the texts than at the paper EPT.	1	2	3	4	5
14	I spent more time reviewing the texts than at the paper EPT.	1	2	3	4	5
15	At the computer EPT I spent more time correcting spelling and grammatical errors than at the paper EPT.	1	2	3	4	5
16	Do you feel that you are misplaced based on your score for the paper test?	1	2	3	4	5
17	Do you expect you will be placed appropriately on computer test rather than paper test?	1	2	3	4	5
18	If you answer ’YES’ in question 17, why do you feel like that?					

		Strongly disagree			Strongly agree	
Only if you think you wrote better at the computer-delivered EPT, please proceed to the next items. Otherwise please skip to item 25.						
19	I wrote better because the given topic was easier than the one before.	1	2	3	4	5
20	I wrote better because I was able to write more sentences.	1	2	3	4	5
21	I wrote better because I was able to organize better.	1	2	3	4	5
22	I wrote better because I was able to correct more errors.	1	2	3	4	5
23	I wrote better because reviewing the text was easier.	1	2	3	4	5
24	I wrote better because functions in a computer such as 'copy and paste' promote easier text development.	1	2	3	4	5
25	If you have any comments or suggestions regarding the computer EPT, write them below.					

Appendix B. Benchmarks for EPT composition scoring (graduate level)

Too low
- Insufficient length
- Extremely bad grammar
- Doesn't write on assigned topic; doesn't use any information from the sources
- Majority of essay directly copied
- Summary of source content marked by inaccuracies

ESL 400
- May contain an Intro, Body, and Conclusion (generally simplistic)
- Does not flow smoothly; hard to follow ideas
- Lacks development and/or substantial content
- May be off topic
- Little or no use of sources to develop ideas
- Lacks synthesis (of ideas in the two sources or of source and the student's own ideas)
- Summary of source content contains minor inaccuracies (details) and possibly major inaccuracies (concepts)
- Lacks sophistication in linguistic expression; little sentence variety and sentence complexity not mastered

- Lexico-grammatical inaccuracies are frequent and impede comprehension; awkwardness
- Attempts at paraphrase are generally unskillful and inaccurate

ESL 401

- Usually contains an Intro, Body, and Conclusion (reasonable attempt)
- Some development or elaboration of ideas
- Some use of sources to develop ideas
- Writes on topic
- Some synthesis (of ideas in the two sources or of sources and the student's own ideas)
- Summary of source content may contain minor inaccuracies (details)
- Neither simplistic and awkward nor smooth and sophisticated
- Some sentence variety and complexity
- Some grammatical, lexical errors; essay still comprehensible
- Moderately successful paraphrase (in terms of smoothness and accuracy)

Exempt

- Usually contains an Intro, Body and Conclusion
- Substantive content and effective elaboration (whether based on sources or own ideas)
- Writes on topic
- Good synthesis (of ideas in the two sources or of source and the student's own ideas)
- Summary of source content usually accurate
- Effective skillful paraphrase (ideas accurate and smooth expression)
- Smooth flowing (may be sophisticated with lexical and sentence variety and complexity)
- Strong linguistic expression (in terms of grammar, vocabulary, and style)

Appendix C. Feature analysis form

Essay ID: ________ Reader's name: __________

Directions: There are four statements in a feature. Please mark on a statement which you think best represents the feature of the essay.

ORGANIZATION

_____ No organization of ideas, insufficient length to ascertain organization

_____ May contain an Intro, Body &Conclusion (generally simplistic) lack of paragraph and essay cohesion (doe not flow smoothly, hard to follow ideas)

_____ Usually contains an introduction, body, conclusion (reasonable attempt); some paragraph and essay cohesion

_____ Clear plan; excellent introduction, body, conclusion; cohesion at paragraph and essay levels.

CONTENT

_____ Doesn't write on assigned topic

_____ May be off topic, Lacks develop and/or substantial content,

_____ Some development or elaboration of ideas, Some use of sources and develop ideas, Writes on topic

_____ Main idea developed well. Substantive content and effective elaboration

LINGUISTIC EXPRESSION

_____ Extremely bad grammar; totally incomprehensible. No sentence variety or complexity

_____ Grammatical/lexical inaccuracies are frequent and impede understanding; awkwardness. Lacks sophistication in linguistic expression; little sentence variety and sentence complexity not mastered, Little sentence variety and sentence complexity not mastered.

_____ Some grammatical/lexical errors, but still comprehensible; some sentence variety and complexity, Neither simplistic and awkward nor smooth and sophisticated

_____ Strong linguistic expression in terms of grammar, vocabulary and style, May be sophisticated with lexical and sentence variety and complexity

USE OF SOUCES

_____ Doesn't use any information from the sources, Majority of essay directly copied Summary of source content marked by inaccuracies.

_____ Lacks synthesis (of ideas in the two sources or of source and the students' own ideas), Attempts at paraphrase are generally unskillful and inaccurate. Summary of source content contains minor inaccuracies (details) and possibly major inaccuracies (concepts)

_____ Some synthesis (of ideas in two sources or of source and the student's own ideas) Summary of source content may contain minor inaccuracies (details) Moderately successful paraphrase

_____ Good synthesis, Summary of source content usually accurate. Effective skillful paraphrase

LENGTH Positively affected, Not affected, Negatively affected

MECHANICS Positively affected, Not affected, Negatively affected