ACT Research Report Series 93-8

A Comparison of Word-Processed and Handwritten Essays From a Standardized Writing Assessment

Edward Wolfe
Sandra Bolton
Brian Feltovich
Catherine Welch

December 1993

ACT

A Comparison of Word Processed and Handwritten Essays from a Standardized Writing Assessment

Edward W. Wolfe

Sandra Bolton

Brian Feltovich

Catherine Welch

American College Testing

Running head: ESSAY COMPOSITION

Abstract

The two studies described here compare essays composed on word processors to those composed with pen and paper for a standardized writing assessment. The following questions guided these studies: 1) Are there differences in test administration and writing processes associated with handwritten versus word processor writing assessments?, and 2) Are there differences in how raters evaluate handwritten versus word processor format? Study 1 revealed that there are some differences in the manner in which students approach writing essays when given a choice of the two formats. Study 2 revealed that there are differences in the manner in which essays in each format are scored by raters.

Table of Contents

Page

Background 4

Purpose 7

Study 1 8
Design 8
Results 9

Study 2 10
Design 10
Results 12

Discussion 15

References 18

Table 1: Descriptive Statistics for each Mode 20
Table 2: Inter-reader Correlations for each Mode 20
Table 3: Generalizability Study Results 21
Table 4: Mode x Version ANOVA Results 21
Table 5: Mean Frequencies for Processing Option Use 22
Table 6: Mean Frequencies for Content Categories 22
Table 7: Processing Actions for Essay Scoring 23
Table 8: Example Coding Sheet 24

Appendix A 25

A Comparison of Word-Processed and Handwritten Essays from a Standardized Writing Assessment

Background

Recent reforms in writing assessment have called for methods of assessment that are both authentic and direct (Frederiksen & Collins, 1989 and Wiggins, 1989). However, the adoption of essay formats in writing assessment may introduce sources of construct irrelevant variance into test scores that have not typically been considered by test developers. Controlling these sources of measurement error is an important part of insuring the reliability and validity of direct writing assessments. A central issue for establishing the defensibility of new forms of assessment is construct specification. That is, adequately transmitting the scoring criteria from test developers to test scorers and consumers becomes paramount in establishing validity and reliability for essay assessments.

One potential source of construct irrelevant variance that must be taken into account by developers of direct writing assessments is textual appearance (e.g., handwriting quality and response length). Handwriting quality has been acknowledged as a factor that is difficult for raters to ignore. Markham (1976) studied the effect of handwriting quality on grading by asking elementary school teachers and student teachers to score papers of varying content sophistication and handwriting quality. These teachers rated papers with neater handwriting consistently higher than those with poor handwriting regardless of the quality of content. Marshall (1972) performed a similar study with secondary history teachers, asking them to judge the content of essays with varying levels of content sophistication, legibility, and composition errors. Results indicated that composition errors have detrimental effects on the

grades assigned to typed essays, that handwritten essays are assigned lower grades than typed essays free of composition errors, and that composition errors do not have systematic effects on grades assigned to handwritten essays.

These studies imply that handwriting quality has differential effects on the grades assigned to student essays with similar content. However, the causes of this effect were not investigated. One possible cause was suggested by Huck and Bounds (1972). These researchers identified essay readers with varying degrees of handwriting neatness and asked them to score essays with different levels of content sophistication and handwriting quality. Neat writers assigned higher grades to neat essays, but messy writers did not differentiate essays based on handwriting quality. In another study, Chase (1979) asked essay raters with prior knowledge of a group of hypothetical students' achievement to score essays with identical content and varying qualities of handwriting. Raters who had been given "high" expectations graded more liberally than did readers with "low" expectations. This appeared to be especially true when the paper read was in poor handwriting. When writing was less legible, the readers depended more heavily on expectancy, with the high expectancy group getting higher scores. When handwriting was legible, however, the impact of expectancy diminished.

These studies suggest that reader characteristics and beliefs may interact with handwriting quality in essay scoring. Such an effect may be greatly compounded when considering the influence of atypical testing conditions, such as using word processors to compose essays, on essay scoring. Unfortunately, although composing essays on computers is becoming more common, studying its effects on writing assessments has received little attention. Arnold, Legas, Obler, Pacheco, Russell, and Umbdenstock (1990) performed one series of studies of the effects word processing had on essay scores in the context of community college placement examinations. In their first study 300 handwritten essays

(HW-O) were transcribed verbatim to word-processed copies (HW-T) and scored by trained readers. The word processor copies received scores .3 units lower on average, on a six-point scale than the hand written originals.

In a follow-up survey, readers reported preferring HW-O papers even though they were more difficult to read than the HW-T essays. Readers also reported having higher expectations of word-processed papers and empathizing more with the writers of handwritten papers. In a third study, students were surveyed to identify why they chose, given the opportunity, to use either word processors or pen and paper to compose their essays. Students who produced handwritten papers reported feeling uncomfortable about their typing skills, computer experience, or technology in general, and that these problems might effect their test scores. Students who chose to use word processors did so because corrections (e.g., spell-checking) are easier, and they thought the papers would look better. Students in this study chose handwriting over word-processing three to one.

Another set of studies on this problem was performed by Powers, Fowles, Farnum, and Ramsey (1992). Their purpose was to determine the effects of the mode of writing (handwriting or word processor) on essay scores. Sixty-four essays (two each from thirty-two students) were scored on a six-point holistic scale. Students produced one essay on a word processor (WP-O), and the other was originally handwritten (HW-O). In addition, handwritten originals were transcribed verbatim to word processor copies (HW-T), and word processor originals were transcribed to handwritten copies (WP-T) with only obvious typographical errors omitted.

In all cases, papers scored from handwritten originals or transcripts received higher scores. Writing researchers examined the papers and determined that word-processor versions *appeared* to be shorter in length, that poor handwriting often masked mechanical problems that were more apparent in word processing papers, and that handwritten originals

showed more obvious signs of revision than word processor essays. In an attempt to compensate for these problems, reader training was structured to emphasize that handwritten and word-processed papers make different impressions and that appearance of length may be influenced by using a word processor. Papers written in both modes were used during training, and trainers checked for differences in the standards applied to scoring essays in the two modes. Also, word-processed papers were double-spaced to decrease the appearance of length differences. Again, handwritten transcriptions received higher scores than word processor originals. However, these differences were smaller than those observed previously.

These studies provide some interesting insights into the possible effects of word processors on essay scoring. First, the quality of a writer's handwriting influences scores; essays written with poorer quality handwriting receive lower scores. Second, reader beliefs and expectations may influence essay scores with papers that are attributed to higher expectations being critiqued according to more stringent standards. Third, the influence of word processing on essay writing and scoring is not yet clear. It is apparent that word-processed papers are scored more stringently than handwritten ones. However, it is not clear whether there are significant qualitative differences in the manner in which essays are composed in each mode, in the content of resulting essays, or in the methods readers use to score these papers. These issues are addressed in our studies.

Purpose

The purpose of the two studies described here is to compare essays composed on word processors to those composed with pen and paper. The following questions guided these studies: 1) Are there differences in test administration and writing processes associated with handwritten versus word processed writing assessments?, and 2) Are there differences

in how raters evaluate handwritten versus word processed essays? These questions were addressed by the studies described below.

Study 1

Design

Study 1 was designed to determine if there are differences in the manner in which responses to a large-scale standardized writing assessment are composed due to the mode of composition. Subjects (N = 157) for this study were tenth-grade students from three Midwestern high schools chosen to be representative of a variety of socio-economic and cultural backgrounds. The schools were confirmed to have good on-site computing facilities used in teaching writing. Students in each school were administered a standardized writing assessment. About half of the students (N = 80), distributed evenly among the three schools, wrote their responses by hand (HW) and the other half (N = 77) composed essays on word processors (WP).

The writing assessment was identical for both modes of presentation, handwritten and word-processed, with two 30-minute periods for writing. The first period was used to produce a rough draft of a paper and the second period, on the following day, was used to revise and rewrite the draft. On the first day, the students were given a writing prompt and several prewriting activities to help them get started with their drafts. At the end of the period, the students were given some questions to help them think about how they might revise their work. The next day, the students were asked to look back at their rough drafts and the revision questions as well as any notes they had made before writing the final draft of the essay.

In each case, a separate classroom was used for each mode so that distractions would be minimized. The teachers who administered each mode of the assessment read a standardized script that differed only in reference to the mode of writing (e.g., "writing"

versus "typing" or "pen" versus "keyboard"). Students chose one of the two administrative formats and completed the writing assignment on two consecutive days. Drafts and final versions of the writing were collected from each student. The test administration in one of the three schools was observed by an ethnographer. Afterwards, students and teachers were informally interviewed concerning their feelings about the testing process.

The observed computer-equipped classroom had 25 identical computers positioned around the side walls. The classroom also contained rows of desks facing the front of the class in the middle of the room. Students sat at the desks while instructions were read and then moved to the computers to compose their essays. In order to avoid giving handwriting students a time advantage, the machines were already running with the word processor loaded when the students arrived for the assessment. Students had access to a word processor, commercial grammar-checking software, and software that the teacher had written to automatically check for common stylistic faults. There was one printer for every four computers. The setup of the observed handwriting class was the same with respect to arrangement and resources with the exception of the computers.

Results

Several differences between the WP and HW writing processes were observed. The first and most obvious was the WP students' frequent use of the spelling-, grammar-, and style-checking software. Students using the computers were almost unanimous in their enthusiasm for the computer's ability to check their work. Nearly 90% of the computer users ran at least one of these programs prior to printing a rough draft of the essay. Most of the checking done by WP students on the first day was performed on a surface level. Some students ran the style-checking program several times. Some students were observed using the page preview feature of the word-processing program to insure that the output would look polished and professional. However, on the second day, most of the WP students were

observed reading from a printout of the style-checking software while they revised their work, especially in those areas flagged by the style-checking program.

The instructions for the assessment in both classrooms encouraged students to use whatever means they desired to revise their rough drafts. However, none of the HW students were seen using dictionaries to check spelling or asking their classmates or teacher for help or advice regarding grammar or style. Their editing routines included reading the rough draft, marking problems (e.g., mechanical errors), and rewriting short passages.

When interviewed, the HW students seemed less comfortable about their typing abilities. One student, when asked why she had chosen to write her paper rather than to use the computer, replied that she did not like using the computer because "... it tells me how stupid I am."

Other issues became apparent during the observations. One was the extent to which students were able to see each other's work. The narrative prompt elicited writings of a personal nature, encouraging students to share personal stories and emotions. Because the computers were positioned less than a foot apart, several students were observed to surreptitiously read work on the neighboring monitors. Interestingly, and probably related, many WP students used very small fonts; several were so small that they were unreadable from adjacent computers. This sharing of work or observing the work of others was not apparent in the HW classroom.

Study 2

Design

In Study 2, analyses were performed to determine whether or not there were differences in the method used to score word processed and handwritten papers. Each of the 157 papers was assigned a rating by two independent raters who were randomly-selected from a group of 18. Ratings were based on a six-point holistic rating scale. The group of

readers was composed of twelve females and six males. The average age was 34 years. All readers were Caucasian with the exception of one African American. Half of the readers had been teachers within the last three years with most of the teaching experience occurring at the university level. One reader had obtained a high school diploma, seven had received a Bachelor's degree, and ten had received a Master's degrees. These degrees were representative of a variety of major areas of study. Only three readers reported having professional writing experience, and only two reported having more than one year of experience as a professional essay reader.

Four of these readers (two females and two males) were randomly selected to perform a think-aloud task in which three papers (at least one example of a word processed paper and one example of a handwritten paper) were scored as the reader verbalized his or her thoughts. Based on the model of scorer cognition presented by Wolfe and Feltovich (1994), protocols were divided into phrases that contained a complete thought (t-units). Each t-unit was coded according to the *process action* being performed by the reader. For example, prior studies have shown that readers typically use the *reading* process to construct an image of the text written by the student. While reading, the scorer may *monitor* the text image for certain elements of writing and may *comment* about the scoring method being used or the characteristics of the writing. After completing the reading, readers often *review* the contents of the essay or *compare* it to other papers recently read. Finally, the reader *decides* what score to assign and provides a *rationale* for why the paper deserves the assigned score.

Each statement was also coded according to the *content* being cited. It is important to note that content focus is primarily derived from the scoring rubric, and it is defined as the values and parameters upon which scoring decisions are made. For this study, the scoring rubric emphasizes *development* of ideas, *organization* of the writing, the use of a writer's *voice* through sentence structuring and word choice, and control of *mechanics*. Readers may

also make general, *non-specific* comments such as "This is really good." Finally, readers may bring prior values to a scoring session so that other aspects of the essay, such as textual *appearance* or *subject* selection, may be noted during scoring. Appendix A contains a more detailed discussion of the coding system.

Prior to scoring, all handwritten original responses (HW-O) were transcribed verbatim to word processor copies (HW-T) using a variety of font sizes and print qualities (in order to randomize these effects), and all word processor original responses (WP-O) were transcribed verbatim to handwritten copies (WP-T) of varying handwriting quality. Transcriptions were performed by a variety of writers in order to insure a randomness of quality of handwriting. Each of these writing samples was scored by two readers selected at random. Another set of analyses was performed in order to determine the differences between the original and transcribed version of word processor and handwritten responses.

Results

Table 1 shows the descriptive statistics for the four groups of papers scored. This table shows that transcriptions under both modes were scored lower than the originals. The standard deviations for scores are all of similar magnitude.

Insert Table 1 about here

Reliability of reader performance was estimated by computing the interrater correlation for the two independent ratings of each paper. The interrater correlations differentiated the two formats and their transcriptions. Table 2 shows the interrater correlations for each form. The Word Processor originals were rated with an average correlation of $r = 0.76$ while their handwritten transcriptions were rated with an interrater correlation of $r = 0.68$. On the other hand, the interrater correlation of the handwritten

originals was r = 0.64 while their word processor counterparts were rated with r = 0.67. A generalizability study (G-Study) revealed that the proportion of observed variance accounted for individual differences between students was higher when handwritten original essays were transcribed to word processed copies (δ = 0.05). However, the opposite effect was observed when word processor essays were transcribed to handwriting (δ = 0.06), in favor of word processor essays). These results are shown in Table 3.

Insert Table 2 about here

Insert Table 3 about here

An Analysis of Variance showed a significant difference between the scores assigned to originals and transcribed papers (F = 19.42, df = 1, p = 0.015) with originals being scored higher regardless of mode of composition. The mean difference was δ = 0.25. Table 4 contains the results of the ANOVA. The lack of interaction between mode and version indicates that the transcription effect was consistent across in both directions (i.e., word processor papers transcribed to handwriting and hand written essays transcribed to word processor). The correlation of scores between handwritten originals and their transcribed versions was 0.76 and the correlation between word processor originals and their transcribed versions was 0.67 (not comparable) indicating that the two versions were not consistently scored for the same qualities.

Insert Table 4 about here

Content analyses of differences in the original versus the transcribed papers by experts in writing assessment revealed that transcribed papers differed from their originals in five ways. First, there were differences in the apparent length of the transcriptions. Because of line spacing and handwriting size, all original essays (whether handwritten or word processed) appeared longer than their transcriptions. In some cases, originals ran one page longer than the transcribed version. Second, paragraphs in the handwritten original papers seemed longer than in the transcribed versions. This was not as apparent when word processor essays were transcribed to handwriting. Third, transcribers added a number of errors to both types of copies. An average of two errors (spelling or typographical) were added to each paper reviewed. Fourth, transcribing papers from handwriting to word processors made the errors that students committed more noticeable. This effect was not true for transcribing word processor essays to handwriting. Finally, in some instances the handwritten copies of word processor originals looked sloppier written when compared to the handwritten originals.

Analyses of the think-aloud protocols revealed differences in the way that word processor original essays (WP) and handwritten originals (HW) essays were judged. First of all, consistent with the fact that word-processed essays received higher scores is the fact that readers made more positive comments about WP papers ($\delta = 1.83$) and more negative comments about HW papers ($\delta = 1.50$). Second, it seems that these readers used different processes to evaluate the two types of papers. Table 5 shows the mean frequency with which each *process action* was used by the readers during scoring. When reading HW essays, readers tended to read less of the paper at one time, stopping more often to make evaluative comments about the essay. But, when reading word-processed papers, readers interrupted their reading less and saved most of their comments until after completing the entire paper.

Insert Table 5 about here

Also, critiques of the WP papers tended to focus on the *development* of papers (e.g., elaboration and support of ideas, use of narrative elements and figures of speech, etc.) while HW comments focused on the emergence of *organization* and the writer's personal *voice* in the writing. WP papers also received more comments concerning their format (e.g., "I don't like the justification here.") as well as the *subject* upon which they were written (e.g., "This is a rather mundane topic."). Table 6 shows the mean frequency of citations of *content* by these readers. Finally, the nature of non-evaluative comments differed for the two types of papers. WP comments referred to the *method* through which readers planned to arrive at a score ($\delta = 1.00$) while HW comments referred to the *reading* process or characteristics of the writer ($\delta = 1.67$).

Insert Table 6 about here

Discussion

The results of these studies have implications for both practice and research on direct writing assessment.

The method of revising and editing used in each mode of composition was different. Most notable is the fact that students who used word processors were more likely to seek assistance during composition by using spelling and style-checking utilities. However, this difference is also confounded by the self-selection methodological problem mentioned previously. It may be the case that because of different experiences that lead students to

choose word processing over handwriting also caused them to use different writing strategies while composing their essays.

The most interesting finding is that scores assigned to original essays were not equivalent to those assigned to transcribed versions. It should be noted that this finding contradicts a study by Powers, Fowles, Farnum, & Ramsey (1992) that showed handwritten essays to be scored higher whether they were originals or transcribed versions of word processor original essays. Our study showed that transcribed versions received lower scores than originals regardless of the mode of composition. This lowering of scores also had some influence on the ordering of students as reflected by the mid-ranged correlations between original essays and their transcriptions (i.e., around 50% of the variance in original scores may be accounted for by the variance in transcription scores). Other differences that may be more related to the final format may also factor into the lowering of scores on transcribed essays. For example, the apparent length of transcribed papers and paragraphs within those papers was shorter than in the originals or errors may be more apparent depending on the textual format.

One explanation may be that scorers focus on different standards when scoring word processed papers than they use for handwritten essays. The think aloud protocols of our scorers would suggest that scorers may be focusing on more simplistic or concrete features of word processing essays, and that they focus on more conceptual and abstract dimensions of writing when the essay is handwritten. For example, our scorers were more likely to mention compliance with the prompt or the appearance of the text when scoring word processor essays. Conversely, when a handwritten essay was being scored, the scorer was more likely to mention the essay's organization or the emergence of a writer's voice as an evaluative consideration. In support of this conclusion is the fact that non-evaluative comments made while scoring word processor essays focused on the method being used by

the scorer (e.g., "I need to go back and look at the organization of the paper.") rather than on the writer or essay characteristics (e.g., "This writer has done a lot of reading.") as was the case with handwritten essays.

Content analyses of the word processor (WP) and handwritten (HW) originals revealed striking differences in the overall quality of the two sets of essays. WP essays were longer. They contained about 75 more words than did HW essays (t = 3.03, df = 138, p = 0.00). The word processed papers were also of higher quality. The topics in WP papers were wide-ranging and engaging while those in the HW sample were on narrower and more private topics (e.g., accidents, divorce, and death). WP writers also related the most memorable parts of their experience to the reader. The writings in the HW sample tended to be simple and general chronologies. Seventy percent of the WP writers used dialogue. Dialogue appears only once in HW essays. The vocabulary contained in WP writings was also more precise and complex than that found in HW essays. Finally, the HW papers contained almost twice as many mechanical errors (28) as the WP papers (16).

Students in this study were self-selected into the composition mode groups. Therefore, the differences observed between word processor original essays and handwritten originals may be attributable to factors other than the mode of composition. For example, it may be the case that many of the observed differences between the word processor and handwritten originals are due to socio-economic differences in the groups who chose each mode of composition. Students who chose to compose their essays on the word processors demonstrated confidence in their keyboard abilities. This confidence may be the result of educational experiences that correlate with proficiency in writing. Future studies should take these differences into account.

References

Arnold, V., Legas, J., Obler, S., Pacheco, M.A., Russell, C. & Umbdenstock, L. (1990). *Direct writing assessment: A study of bias in scoring hand-written vs. word-processed papers.* Rio Hondo College, Whittier, CA.

Chase, C.I. (1979). The impact of achievement expectations and handwriting quality on scoring essay tests. *Journal of Educational Measurement, 16(1),* 39-42.

Ericsson, K.A. & Simon, H.A. (1984). *Protocol analysis.* Cambridge, MA: MIT.

Frederiksen, J.R. (1992). *Learning to "see": Scoring video portfolios or "beyond the hunter-gatherer in performance assessment".* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Frederiksen, J. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18(9),* 27-32.

Huck, S.W. & Bounds, W.G. (1972). Essay grades: An interaction between graders' handwriting clarity and the neatness of examination papers. *American Educational Research Journal, 9(2),* 279-283.

Markham, L.R. (1976). Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal, 13(4),* 277-283.

Marshall, J.C. (1972). Writing neatness, composition errors, and essay grades reexamined. *The Journal of Educational Research, 65(5),* 213-215.

Powers, D.E., Fowles, M.E., Farnum, M., & Ramsey, P. (1992). *Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays.* RR. 92-45. Educational Testing Service: Princeton, NJ.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70(9),* 703-713.

Wolfe, E.W. & Feltovich, B. (1994). *Learning how to rate essays: A study of scorer cognition.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Table 1: Descriptive Statistics for each Mode

Mode of Composition	*Version of Essay Scored*	
	Original	Transcribed
HW	N = 80 Mean = 3.50 SD = 0.90	N = 80 Mean = 3.27 SD = 0.78
WP	N = 77 Mean = 4.10 SD = 1.02	N = 77 Mean = 3.83 SD = 0.90

Table 2: Inter-reader Correlations for each Mode

Mode-Version	*Inter-reader Correlation*
HW-O	0.64
HW-T	0.67
WP-O	0.76
WP-T	0.68

Table 3: Generalizability Study Results

Mode-Version	*Source*	*Variance Component*	*G Coefficient*
HW-O	Student	0.6287	0.62
	Rater	0.0398	
	Error	0.3477	
HW-T	Student	0.4781	0.67
	Rater	0.0396	
	Error	0.1916	
WP-O	Student	0.9137	0.74
	Rater	0.2235	
	Error	0.0947	
WP-T	Student	0.6585	0.68
	Rater	0.0078	
	Error	0.3039	

Table 4: Mode x Version ANOVA Results

Source	**SS**	**DF**	**MS**	**F**	**p**
Mode (HW/WP)	105.559	1	105.559	32.554	0.000
Version (O/T)	10.421	1	19.421	5.989	0.015
Mode x Version	0.096	1	0.096	0.030	0.863
Error	1005.19	310	3.243		

Table 5: Mean Frequencies for Processing Option Use

Mode	Read	+ : - Comments	Decide	Monitor	Review	Compare	Diagnose	Rationale
WP	3.83	5.00 : 2.83	1.67	2.17	3.17	1.00	0.83	2.67
HW	5.17	3.17 : 4.33	2.00	3.00	3.00	1.33	0.50	1.33

Table 6: Mean Frequencies for Content Categories

Mode	Appearance	Development	Grammar	Non-Specific	Organization	Subject	Voice
WP	0.50	2.83	1.00	1.17	1.83	0.33	1.17
HW	0.17	1.33	1.33	0.83	2.33	0.00	2.00

Table 7: Processing Actions for Essay Scoring

Class	Action	Definition	Associated Knowledge
Interpretive		Actions used to create a text image or to clarify points of consideration	--
	Read	Read from the student response to create a text image	Text
Evaluative		Actions used to map the model of performance onto the text image	--
	Decision	Declare a score or range of scores for a given response	*Valence*
	Monitor	Reference elements of the text or text image in terms of the reader's model of performance during reading (i.e., making notes)	*Content & Valence*
	Review	Reference elements of the text or text image in terms of a reader's model of performance after completing the reading (i.e., taking stock)	*Content & Valence*
Justification		Actions used to check the accuracy of a decision or to provide a rationale for a given decision	--
	Compare	Comparing elements of the text or text image to some other source of knowledge	*Source*
	Diagnose	Describe the shortcomings of the paper or how it could be improved	*Content & Valence*
	Rationale	Reference elements of the text or text image in terms of reader's model of performance that are used as support for a given decision	*Content & Valence*
Interactive		Actions that are used to provide peripheral information about the rating experience	--
	Comment	Provide information about a number of parameters of the rating experience	*Parameter*

Table 8: Example Coding Sheet

Action	Source/Content/ Parameter	Valence	Comment
Read	Words 1 - 141		
Monitor	Development	N/F	"Uses figurative language and ellipses unsuccessfully"
Read	Words 142-430		
Comment	Reading		"I have to watch my prejudices against religious papers"
Comment	Reading		"I have trouble with papers that use figurative language that gets out of control"
Review	Development	N/F	"Want to give credit for using the metaphor"
Review	Organization	N/F	"Not a paragraph paper"
Review	Organization	+	"Break where there is a good transition"
Review	Mechanics	-	"but not mechanically"
Decision		4	
Rationale	Non-Specific	-	"More out of control"
Compare	Prior		"than other 4's"
Compare	Prior		"but attempts things that a 5 or 6 does"

Action

Interpret: Read
Evaluate: Decision (V), Monitor (C,V), Review (C,V)
Justify: Compare (S), Diagnose (C,V), Rationale (C,V)
Interact: Comment (P)

Source

Prior
Reader
Rubric

Parameter

Scoring
Reading

Valence

Positive (+)
Neutral/Fail (N/F)
Negative (-)
Range (1-6)

Content

Appearance
Non-Specific
Voice
Development
Organization
Mechanics
Subject

Reader ID: CM

Paper ID: 1036896

Appendix A

CODING SYSTEM FOR READER THINK ALOUD PROTOCOLS

The Model of Scorer Cognition

The model of scorer cognition described earlier in this paper is a conceptual map/information-processing model of an essay reader's decision making process. In order to document the components of this model, a think aloud activity is used with essay readers as they score a number of essays. It is assumed that the utterances produced by a scorer engaged in a think aloud task are partial traces of the representations and processes that are executed as decisions about how to rate a particular response are made (Ericsson & Simon, 1984). That is, the method assumes that each statement indicates that a specific processing action has been taken and that that action takes place by manipulating knowledge that is relevant to the decision making process.

The Coding System

The coding system described here was created for analyzing think aloud protocols from essay readers. In this case, thought-units (t-units) from a think aloud protocol can be coded with respect to a number of dimensions. For example, an utterance will indicate that a specific *action* is being taken and that that action is based upon a certain type of knowledge or information (e.g., a certain *content* or criteria classification, a certain *source* of knowledge, or a certain *parameter* of the rating situation). Furthermore, some actions may be judgmental in nature and thus may be related to the assigning of a value judgment (or *valence*) to the essay. The sections that follow further define the range of actions, sources, content, type, and valence that may be observed in think aloud protocols from essay scoring sessions.

Actions

Every statement made by a reader may be coded according to the action being executed. An *action* refers to one of several processes that a reader may perform when making a scoring decision. A processing action is a description of the manner in which a piece of knowledge is manipulated during the scoring of a paper. Processing actions may be classified as being *Interpretive* (those having to do with obtaining information), *Evaluative* (those having to do with the forming of a decision), *Justification* (those having to do with providing a rationale for a decision), or *Interactive* (those having to do with personal insights about the rating and reading task). Table 7 shows the classifications of actions and the specific actions associated with these classes as well as the types of knowledge that may be associated with each action.

Insert Table 7 about here

Content

Content plays an important role in the decision making of an essay reader. *Content* refers to the language and values contained in the reader's model of performance that is used as the "rules" for making scoring decisions. The reader's model of performance is called upon to supply information when a reader executes the following actions: *Monitor, Review, Diagnose,* and *Rationale*. Each of these actions is performed by making a comparison between the text or text image and the contents of the reader's model of performance.

For our purposes, the following content sources {taken from the scoring rubric and pilot studies of scorer cognition may be considered by a rater: the physical *appearance* of the text; the *development* of the writing, *mechanics*; *non-specific* or general comments about writing quality; the *organization* and structure of the writing; *subject* of the essay; and the revelation of insight and use of a personal style, often referred to as *voice*, in the writing. Definitions and examples of statements indicative of each of these content classifications follow.

Appearance: Indications of the quality of the writing or typing contained in a response (including typographical errors or length of response).

- I like the fact that it is typed.
- It is almost unreadable.
- I try to ignore penmanship.
- This paper is of average length.

Development: Development refers to the level of sophistication in using writing to communicate. It includes Details, Elements, and Story. Details refer to the amount of, specificity of, and quality of the information included in a story. It may be called elaboration, development, or support of ideas. Elements refers to one's ability to use elements of writing in communicating the story. It may be called dialogue, character, or setting; as well as control of language. Story refers to one's ability to tell a story. It includes communication ability, interest level, and sophistication of thought & ideas (including the main idea).

Details:
- The writer provides no support for the ideas.
- The paper lacks elaboration
- Few and sometimes no details are given.
- The writer doesn't give me enough information.

Elements:
- The use of dialogue spices the narrative.
- Narrative devices are attempted but aren't always successful.
- The writer lacks control of the story elements.
- The writer attempts to use a metaphor here.

Story:
- The story is easily understood.
- The ideas presented are not very sophisticated.
- The writer achieves her goal.
- The story is very interesting.

Mechanics: Mechanics refers to aspects of the writing that focus on the correctness of form at the word level. It includes Spelling, Punctuation, Grammar, and Usage. Spelling and punctuation refer to the correctness and usage of these elements of writing. Grammar and usage refer to the quality and appropriateness of language usage, grammatic rules, agreement, and syntax.

Spelling & Punctuation:
- "Their" is misspelled.
- I don't like the way the semi-colon is used here.
- There are a few minor mechanical errors.
- The punctuation was fine.

Grammar & Usage:
- Often the language used causes confusion and/or incoherence.
- There are many problems with verb tense agreement.
- The usage and flow of language is smooth.
- This sentence is grammatically incorrect.

Non-Specific: These are general comments about the writing without referring to a specific aspect of the Content itself.

- This is good writing.
- I like it.
- That's good.
- Hmm. Interesting.

Organization: Indications of the quality and clarity of the sequencing, structure and flow of events, and transitions in a story. (includes focus of writing, introductions and conclusions, paragraphing, and rambling)

- The events of the experience do not flow clearly.
- Level one papers have no direction.
- The story rambles.
- The paragraphing seems artificial.

Subject: Subject refers to aspects of the writing that focus on the prompt and the topic for which the writing was composed. Prompt refers to the extent to which a response addresses the requirements of a given prompt. It may be called the content, process, or goal of the writing or as its appropriateness for the audience. Topic refers to how a chosen topic or subject matter influences the quality of a piece of writing.

Prompt:
- Hardly any effort at all.
- The writer made an attempt to tell a story.
- The writer doesn't really ever tell me how he changed (when the prompt asked for this information).
- I think this paper was written about a different prompt.

Topic:
- I don't like "religious" papers.
- The paper is about a rather boring topic.
- This was a good subject for the assignment.
- Level 5 papers are often about rather mundane experiences.

Voice: Indications of the effectiveness of a writer's style and conveying of emotions in a story as well as insight, humor, or reflection. May include reference to sentences and vocabulary. Sentences refers to the quality and complexity or organization of sentence structure in a story. Vocabulary refers to the quality of word choice or vocabulary in a story.

Voice
- The writer is able to stand back and comment--to take a wider look.
- This writer has a limited ability to express emotions.
- I see a lot of thought and insight in this paper.
- I really like the use of humor here.

Sentence
- This paper has poor sentence structure.
- That's an awkward sentence.
- Good sentence complexity.
- Most of the sentences are rather simple.

Vocabulary
- The writer used a lot of 50-cent words.
- The words fit to the story situation.
- Interesting choice of words.
- The vocabulary used was rather limited.

Valence:

Reader comments that focus on **Content** not only identify which elements of the model of performance are being considered, but they also are typically value-laden. Frederiksen (1992) referred to the value assigned to the judgment as *valence*. The valence of an evaluative comment may be *positive* (successful), *negative* (non-successful), *neutral/failed* (indicating average or no value, both positive and negative qualities, or attempted but was not successful). In this coding system these valences are indicated with a plus (+) for positive, a minus (-) for negative, the letters N/F for neutral/failed.

Source

The *Compare* processing action is performed by manipulating some external form of knowledge. In order to do these manipulations, some medium for storing the knowledge is accessed. These mediums may include: 1) *Prior* (paper is compared to other papers that were previously read), 2) *Reader* (scoring of the paper is compared to scores that might be assigned by other readers) 3) *Rubric* (paper is compared to descriptions provided in the rubric).

Parameter

Interactive processing actions (i.e., *Comments*) are performed by relaying information that is not specific to the rating process. A reader may make a comment about the strategy used to arrive at a score. Readers may indicate that they have some type of a personal reaction to the writing. They may also indicate some observation about the writer or the text that does not directly relate to the scoring task. There are two general *parameters* to which reader's *comments* may refer: 1) *Scoring* (those having to do with the criteria being used or

those dealing with the method through which a score is assigned to a paper), and 2) *Reading* (those having to do with personal reactions to the reading or acknowledgement of biases the reader has and those dealing with the text and/or writer of the essay).

An Example:

The following condensed think aloud has been coded as an example of the application of this coding system. The coding sheet is provided in Table 8.

The reader reads 141 words from the essay.

The reader states, "At this point of time I'm seeing a lot of effort on the writer's part to explain himself in figurative language--not always successful. It is a good sign for me that a writer is trying to do more. And the first sentence told me that when he used ellipses."

The reader reads the remaining 289 words in the essay.

The reader states, "Somebody more mature could rate this better than I could, but I immediately have to watch my prejudice ... (*because*) it's a religious paper. ... I also have trouble with writers who use figurative language when it gets out of control. I tend to spend more time scoring them. ... I want to give her credit ... for the way she employs a metaphor. ... It's not a paragraph paper. ... There is a break about two-thirds the way through where it seems the transition is really well-written, but not mechanically."

The reader gives a score. "I'm going to give it a 4 ..."

The reader continues, " ... because it seems more out of control than the usual 4, but is attempting some things that a 5 or a 6 attempts."

Insert Table 8 about here