DOCUMENT RESUME

ED 399 286 TM 025 513

AUTHOR Wolfe, Edward W.; Kao, Chi-Wen
TITLE Expert/Novice Differences in the Focus and Procedures Used by Essay Scorers.
PUB DATE Apr 96
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Essay Tests; Experience; *Holistic Approach; Judges; *Performance Based Assessment; Protocol Analysis; *Scoring; Test Use; *Writing Tests
IDENTIFIERS *Expert Novice Problem Solving; Experts; Large Scale Assessment; *Variability

ABSTRACT

The amount of variability contributed to large-scale performance assessment scores by raters is a constant concern for those who wish to use results from these assessments for educational decisions. This study approaches the problem by examining the behaviors of essay scorers who demonstrate different levels of proficiency with a holistic scoring rubric. Scorers (n=36) for a large-scale writing assessment scoring project performed a think-aloud task as they scored 24 essays. The protocols were analyzed by examining differences in the ways expert, intermediate, and novice scorers processed each essay and the content focused on in each essay. Two conclusions are drawn, based on these analyses. The first is that expert scorers are more likely to use more fluent processing methods to score essays than are nonexperts. The second is that there is little evidence to suggest that the content that is focused upon by essay scorers is related to scoring proficiency. (Contains 2 figures, 6 tables, and 17 references.) (Author)

Running head: Essay Scoring Expertise

Expert/Novice Differences in the Focus and Procedures Used by Essay Scorers

Edward W. Wolfe

American College Testing, Iowa City, Iowa

Chi-Wen Kao

University of Virginia

Paper presented at the Annual Meeting of the American Educational Research Association in New York, NY (April, 1996).

Abstract

The amount of variability contributed to large-scale performance assessment scores by raters is a constant concern for those who wish to use results from these assessments for educational decisions. This study approaches this problem by examining the behaviors of essay scorers who demonstrate different levels of proficiency with a holistic scoring rubric. Scorers (N = 36) for a large-scale writing assessment scoring project performed a think aloud task as they scored 24 essays. The protocols were analyzed by examining differences in the ways expert, intermediate, and novice scorers *processed* each essay and the *content* of the essay that was focused upon. Two conclusions are drawn, based on these analyses. *Conclusion 1: Expert scorers are more likely to use more fluent processing methods to score essays than are non-experts. Conclusion 2: There is little evidence to suggest that the content that is focused upon by essay scorers is related to scoring proficiency.*

Expert/Novice Differences in the Focus and Procedures Used by Essay Scorers

Achieving levels of reliability that allow large-scale direct writing assessment results to be used for evaluation and selection decisions is a major hurdle for test developers. Generalizability studies have shown that one of the most influential sources of measurement error associated with essay scores is instability of performance across performance tasks (Linn, 1993). Other sources of construct irrelevant variance include rater effects, rater by prompt interactions, and rater by student interactions. Unfortunately, the causes of rater variability are not well understood.

The purpose of the following study is to identify characteristics that differentiate essay scorers of different levels of competence. Although the literature on this topic is not conclusive about why some scorers are more consistent than others, it does offer insights into variables that might account for individual differences in this domain of performance. By identifying the characteristics that differentiate better scorers from poorer ones, scoring trainers could use these characteristics as a means for identifying individuals who demonstrate a better potential to become competent scorers prior to training or to incorporate steps to foster the development of these characteristics in writing assessment training sessions.

Theoretical Background

Most cognitive theories maintain that individuals construct mental representations of objects and processes to understand the world around them. These constructed representations serve as frameworks for interpreting stimuli from the environment (i.e., interpretive frameworks). In recent work in the area of performance assessment, Frederiksen (1992, April) suggested that teacher evaluators use interpretive frameworks to understand and evaluate teacher performance. Frederiksen's notion of an interpretive framework suggests that

the evaluator monitors a performance for some set of criteria (defined by the evaluator's interpretive framework). When a noteworthy instance of a performance criteria occurs, the evaluator makes a mental note of the aspect of the criteria being demonstrated and the degree of competence shown at that instance. Thus, the interpretive framework serves as a means for understanding and recognizing the parameters of the performance being assessed. After all noteworthy moments have been observed, the evaluator considers all of the observations, weights them, and decides on a score. The final step in the process is to create a rationale. Thus, the interpretive framework is also used to organize and communicate ideas about a teacher's performance to others.

A similar process may be used by essay scorers as they evaluate writing quality. Freedman and Calfee (1983) describe an information-processing model of essay scoring that identifies three processes that are essential to evaluating a composition: 1) *reading text to build a text image*, 2) *evaluating the text image* and 3) *articulating the evaluation.* Each of these processes is affected by characteristics of the scorer (e.g., reading ability, world knowledge, expectations, values, and productive ability) and the environment (e.g., time of day, length of task, type of text, the physical environment, the kind of training and supervision, the purpose of the assessment, and the intended audience of the scores). Figure 1 shows how these factors interact.

Insert Figure 1

In the Freedman and Calfee model, information is taken from the printed text, and an image of the student's writing is constructed. The scorer *interprets* writing based on his or her own world knowledge, beliefs and values, and knowledge of the writing process. Aspects of the reading environment may also influence the form that this text image takes. This

means that the text image is not an exact replication of the original text and that one scorer's text image may be very different than text images constructed by other scorers. Based on the text image, the scorer compares various aspects of the writing to representations of the scoring criteria. Through this process judgments are made about the text, and a decision is formulated about how well the writer has demonstrated competence in writing. Finally, the evaluative decision is articulated through written or oral comments about the text.

A comparison of the models of Frederiksen (1992, April) and Freedman and Calfee (1983) suggests two very different approaches to scoring. The Frederiksen model describes an iterative approach to scoring in which a scorer makes multiple evaluative decisions, each successive decision being a revision of the prior one based on additional information obtained from reviewing the material. The Freedman and Calfee model, on the other hand, describes a linear approach in which the scorer creates a holistic image of the text and arrives at a scoring decision by comparing the text image to a mental representation of the scoring rubric. Both of these models acknowledge that different scorers may base their scoring decisions on different features of the performance. However, neither of these models acknowledges that different scorers may use different procedures to make the scoring decision. Other researchers (Huot, 1993; Vaughan, 1991) have indicated that there may be a number of variations of these two general approaches to scoring.

Wolfe and Feltovich (1994, April) and Wolfe (1995) extended this work by proposing a model of scorer cognition that allows for both variations in the features of an essay upon which scoring decisions are based as well as variations in the procedures that are used to make these decisions. Their model portrays essay scoring as an interplay of two cognitive features: knowledge representations and processing actions. Knowledge representations are

classified as being either text images, frameworks of writing, or frameworks of scoring. A *text image* is a mental representation of an essay that is created as the scorer reads and interprets the essay. As suggested by Freedman and Calfee (1983), the text image for a particular essay that is created by one scorer may be very different from the text image created by another scorer because of differences in reading skill, background knowledge, or the physical environment in which scoring takes place. A *framework of writing* is a mental representation of the scoring criteria. These representations may also differ from one scorer to another because of differences in scoring experience, values, education, and familiarity with the scoring rubric (Pula & Huot, 1993). A *framework of scoring*, on the other hand, is a mental representation of the process through which a text image is created and subsequently mapped onto a scorer's framework of writing. The framework of scoring serves as a script, specifying how a variety of possible mental procedures, called *processing actions*, are used to read the essay and evaluate it.

Figure 2 depicts how these components work together in the scoring process. It shows that the text is used by the processing actions to create a text image. The image of the text is not a direct replication of the text because different scorers read the text in different reading environments, have different reading skills, and bring different kinds of experiences and knowledge to the scoring task. After the text image has been created, the processing actions, which are executed according to the script specified by the framework of scoring, map the components of the test image onto the framework of writing. From this mapped image, an evaluative decision is made which is then justified. Because the frameworks of writing and frameworks of scoring used by different scorers may not be identical, scorers will come to different scoring decisions for the same essay. This model of scoring cognition was adopted

for the study reported in this paper, so the following sections provide a detailed description of how frameworks of writing, frameworks of scoring, and processing actions are manifested in the behaviors of essay scorers.

Insert Figure 2

Framework of Writing

A *framework of writing* is a mental representation of the characteristics that constitute proficient or non-proficient writing. This is similar in meaning to the term interpretive framework as used in the work of Frederiksen, Sipusic, Gamoran, and Wolfe (1992), who suggest that the most typical sources of information for creating a framework of writing are scoring rubrics, exemplar libraries, other scorers, or test developers. One would expect the framework of writing that is constructed by a proficient scorer to be very similar to the adopted scoring rubric. However, because of a scorer's prior values, the framework of writing used by a particular scorer may not be adequately described by the scoring rubric.

To explore how a scorer's framework of writing is applied to essays in a large-scale scoring project, Wolfe and Feltovich (1994, April) engaged six scorers in a think aloud task on a set of six narrative essays. The scorers were instructed to read each paper aloud and to verbalize any thoughts they had while scoring the paper. Typically, scorers read the paper verbatim, occasionally interjecting comments about the quality of the essay like "'Their' is misspelled" or "I like that metaphor." After reading the paper, the scorer typically announced what score he or she would assign to the paper and provided a rationale for that decision.

The interjections and justifications supplied by these scorers were placed into one of the following categories: 1) *Appearance* (comments referencing the legibility or length of the essay), 2) *Assignment* (the extent to which the student complies with the writing prompt), 3)

Mechanics (ability to control spelling, punctuation, and grammar in writing), 4) *Non-Specific* (general comments about the control or skill demonstrated by the writing), 5) *Organization* (ability to control the structure and focus of the writing), 6) *Storytelling* (ability to communicate ideas in writing, to develop these ideas, to use narrative elements, and to construct a story), and 7) *Style* (ability within a piece of writing to use words and sentences effectively to convey a personal voice). Taken together, these seven categories define a space of features that essay scorers might include in their frameworks of writing (Wolfe, 1995). These categories can be used to identify the content of an essay upon which the scorer focuses (i.e., *content focus*) and may be useful for identifying differences in the underlying frameworks of writing of essay scorers. Table 1 summarizes these content focus categories, which were used as a system for coding the data that were collected in the study reported in this paper.

Insert Table 1

Framework of Scoring

A *framework of scoring* is a mental representation of the process through which a scorer identifies and interprets evidence from an essay and derives a score based on this evidence. It describes the manner in which a scorer processes or manipulates a variety of knowledge representations during the decision-making process. In essence, a framework of scoring serves as a script that the scorer uses to insure fairness and accuracy in grading.

A framework of scoring must contain a number of elements. First, it must represent the *process of interpretation*. This aspect of the framework represents how a scorer takes in information and determines what aspects of the response will be considered as evidence for competence or non-competence. Second, the framework of scoring must contain an

understanding of the *process of evaluation*. This aspect of the framework contains information concerning how discrepancies or inconsistencies in the evidence will be dealt with or how different aspects of the response will be weighted in the decision-making process. Third, the framework of scoring must contain an understanding of the *process of justification*. Because the scoring process is complex and mentally taxing, it is necessary for scorers to learn how to monitor their own performance and attention and how to incorporate corrective feedback into their scoring activities.

Taken together, the processes of interpretation, evaluation, and justification make up the bulk of the framework of scoring used by the scorer, and these processes roughly correspond to the three processes identified by Freedman and Calfee (1983) (i.e., build a text image, evaluate the text image, and articulate the judgment). These three general scoring strategies seem to be used by most essay scorers (Wolfe & Feltovich, 1994; Wolfe, 1995). However, each of these strategies is likely to be performed through the execution of a number of processes. As a result of combining such processes, the manner in which different scorers execute the scripts specified by their frameworks of scoring may appear to be very different. To differentiate these general strategies from the actual processes through which these scripts are enacted, the Wolfe and Feltovich (1994) model uses the term *processing action*.

Processing Action

A *processing action* is one of several mental activities that a scorer may perform when making a scoring judgment. Processing actions are used in combination to complete the script specified by the framework of scoring. Again, evidence for the existence of a variety of types of processing actions is derived from observations of scorers in the pilot study previously described (Wolfe & Feltovich, 1994, April). In this study, not only were

differences in content focus identified (e.g., story, organization, mechanics, etc.), but different raters also utilized these knowledge representations in different ways. For example, some scorers tended to read the entire text from beginning to end, seldom stopping to note its merits or deficiencies. These scorers often reviewed the text afterwards, citing content that subsequently factored into their scoring decision. Other raters stopped quite often while reading to identify strengths and weaknesses in the essay and would occasionally state how these elements were helping them formulate an on-line decision before finishing the entire essay.

These observations led to the identification of a number of processing actions that scorers perform during the decision-making process. Wolfe and Feltovich (1994, April) observed only two processing action during the *Interpretive* phase of the framework of scoring: *read* (read text to create a text image) and *comment* (verbalize personal reactions to the text). During the *Evaluative* phase of the framework of scoring, scorers could *monitor* while reading (make notes of how the text or text image maps onto the framework of writing), *review* essay contents after reading (take stock of how the text or text image maps onto the framework of scoring), or make a *decision* (assign a score or score range). For the *Justification* phase, raters would often *diagnose* ways the essay could be improved, provide a *rationale* by describing how the text image exemplifies certain aspects of the framework of writing, or *compare* elements of the text image to other sources of information. These processing action categories may provide insights into the individual differences between essays scorers as they read an essay and formulate a scoring decision. Table 2 summarizes these processing action categories. These categories are also used (as are the content focus categories) as a coding system in the study reported in this paper.

Insert Table 2

Expertise

Although few studies of holistic scoring have focused on a cognitive model such as the one presented in Figure 2, prior research suggests predictions for differences in essay scorers. There is considerable evidence that the thinking of essay scorers is driven by specialized knowledge structures (i.e., frameworks of writing) and that the primary focus of these frameworks is on the development and organization of an essay (Diederich, French & Carlton, 1961; Freedman, 1979; Huot, 1993; Pula & Huot, 1993; Vaughan, 1991; Wolfe & Feltovich, 1994, April). This finding is consistent with expert/novice studies in other domains that show experts using highly specialized knowledge structures when thinking about their domains of expertise (Chi & Glaser, 1988). Furthermore, individuals with different levels of experience and expertise may emphasize different aspects of the framework of writing (Huot, 1993). This is probably due to their realization of overriding principles that guide the thinking of experts (Chi, Feltovich & Glaser, 1981).

With respect to differences in the use of frameworks of scoring, evidence supports the notion that a variety of scoring strategies may exist (Vaughan, 1991) and that experienced scorers may be more efficient in using these strategies than novices (Huot, 1993). This efficiency may come from the expert's ability to generalize principles within the domain (Berliner, 1986) and the ability to automate relevant thinking patterns through repeated practice (Chi & Glaser, 1988). These characteristics help reduce expert scorers' cognitive processing loads, allowing them to engage in other types of processing like comprehending difficult text passages or diagnosing student writing difficulties (Huot, 1993). Wolfe and

Feltovich (1994, April) found that processing action use may indicate a number of different scoring styles that may be an important consideration for determining scoring competence.

Findings from this line of research are based on only a few studies and are not conclusive about the nature and meaningfulness of individual differences in scorers' frameworks of writing, frameworks of scoring, and use of processing actions. Further investigations into these differences may reveal ways that these variables can be used to assess the efficiency of various reading strategies and the influences that different training approaches for holistic scoring have on the performance of essay scorers.

Method

Hypotheses

Two hypotheses were investigated for this study. *Hypothesis 1* predicted that scorers with different levels of scoring proficiency employ the processing actions (i.e., *monitor, review, diagnose,* and *rationale*) with different emphasis. *Hypothesis 2* predicted that scorers with different levels of scoring proficiency employ the content focus categories (i.e., *appearance, assignment, mechanics, non-specific, organization, storytelling,* and *style*) with different emphasis.

Subjects

Subjects for this study were 36 essay scorers (selected from a pool of 60) who took part in a large-scale writing assessment scoring project. During the scoring project, each subject scored over 200 essays, each essay also scored by a randomly-selected second reader. An intraclass correlation (r_{ic}) was computed for each subject by comparing the scores assigned by that subject to the scores assigned by the randomly-selected second scorer. This intraclass correlation was used as an indicator of each rater's level of interrater agreement.

Subjects for this study were selected to equally represent three scoring proficiency groups (12 per group): *novices* (low levels of demonstrated interrater agreement, average r_{ic} = .74), *intermediates* (moderate levels of demonstrated interrater agreement, average r_{ic} = .80), and *experts* (high levels of demonstrated interrater agreement, average r_{ic} = .87).

Data

During the scoring project, each subject performed a think aloud task during an interview session. The think aloud task required subjects to verbalize their thinking as they scored 24 essays. Each interview was tape recorded and then was transcribed for analysis. Each complete statement made by a scorer during the think aloud task was coded according to its content focus (i.e., *appearance, assignment, mechanics, organization, storytelling,* or *style*) and its processing action (i.e., *diagnose, monitor, review,* or *rationale*). These data served as a means for comparing experts, intermediates, and novices in the study described here.

Analyses

The numbers of statements made by scorers in each proficiency group that were coded into each content focus category and each processing action category were transformed to proportions to reduce verbosity effects. These proportions were analyzed to determine if the three proficiency groups differ in their focus of attention and method of processing essays as they score. Two *a priori*, orthogonal, two-tailed contrasts were employed for each of the coding categories assumed under each hypothesis. The first contrast compared experts to intermediates and novices [$\Psi_1 = \mu_{experts} - 1/2(\mu_{intermediates} + \mu_{novices})$], and the second contrast compared intermediates to novices ($\Psi_2 = \mu_{intermediates} - \mu_{novices}$). Each pair of contrasts constituted a family of comparisons, and the family-wise error rate was set at α_{FW} = .05 using

the Sidak method to distribute α across the pair of contrasts. The four pairs of *a priori* comparisons resulted in an experiment-wise error rate of α_{EW} =.18 for the processing action analyses. The seven pairs of *a priori* comparisons for the content focus data resulted in an experiment-wise error rate of α_{EW} =.30. For variables that were found to have equal variances (Kohr and Games, 1977) *t* tests were applied to the pair of contrasts. Welch's *t'* statistic was applied to comparisons for which the variances were found to be unequal.

Results

Processing Action Comparisons

Table 3 shows the mean (M_{prop}) and the standard deviation (SD_{prop}) of the processing action proportions for each proficiency group. Table 4 shows the group proportion comparisons with the relevant contrast, degrees of freedom (ν or ν'), the *t* statistic, the associated *p*-value, and the method used to obtain the *t* statistic (*t* or Welch's *t*). Variables for which the variances were statistically different (*monitor* Ψ_1, *review* Ψ_2, and *rationale* Ψ_2) were adjusted using Welch's *t'* statistic. Altogether, eight *t* tests were performed, and two of the comparisons were statistically significant. Experts were less likely to use *monitor* actions than were intermediates and novices; t (21) = -4.45, p = .001. On the other hand, experts were more likely to use *review* actions in their protocols than were intermediate and novice scorers; t (33) = 3.23, p = .005. Statistical tests did not indicate any differences between intermediates and novices for either of these variables.

Insert Table 3

Insert Table 4

Content Focus Comparisons

Table 5 shows the mean (M_{prop}) and the standard deviation (SD_{prop}) of the content focus proportions for each proficiency group. Table 6 shows the group proportion comparisons with the relevant contrast, degrees of freedom (v or v'), the t statistic, the associated p-value, and the method used to obtain the t statistic (t or Welch's t). Variables for which the variances were statistically different, (*organization* Ψ_2 and *storytelling* Ψ_1) were adjusted using Welch's t' statistic. Altogether, 14 t tests were performed, and two of these comparisons were statistically significant. The first statistically significant difference revealed that intermediate scorers were less likely to make *storytelling* comments than were novices; t (33) = -2.74, p = .01. Experts were similar to intermediates in their use of this content focus category as is apparent from the means shown in Table 5. The second statistically significant difference revealed that intermediates were more likely to make *organization* content statements than were novices; t (16) = 2.60, p = .01. However, an inspection of the three group means renders this finding uninterpretable from a perspective that depicts learning as cumulative. That is, the relationship between scoring proficiency and tendency to make comments about an essay's *organization* is not monotonic.

Insert Table 5

Insert Table 6

Discussion

Hypothesis 1 predicted that the proficiency groups would apply the processing actions to essays with different likelihoods. This set of hypotheses was tested by comparing the group proportions of processing action codes for the think aloud protocols. These comparisons revealed that experts are more likely to use *reviewing* behaviors when scoring an

essay. That is, experts were more likely to read the entire essay and then begin to evaluate it. The non-experts, on the other hand, are more likely to break the evaluation process into a series of small evaluation tasks by using *monitoring* processing actions.

Hypothesis 2 predicted that the proficiency groups would apply the content focus categories to essays with different likelihoods. This set of hypotheses was tested by comparing the proportions of content focus categories for the think aloud protocols. Experts and intermediates were found to be less likely to make *storytelling* comments in their think aloud protocols. This is due, primarily, to the tendency of novices to mention the extent to which a writer has produced a convincing and interesting story. A statistically significant difference was also observed between intermediates and novices on *organization*. Novices were shown to be less likely to cite organizational features of the essay. However, because intermediates have a large proportion of *organization* citations, relative to both experts and novices, this finding is not interpretable in a learning context.

These results imply two conclusions about the relationship between essay scoring proficiency and scorer behaviors.

Conclusion 1: Expert scorers are more likely to use more fluent methods of scoring essays than are non-experts. That is, results from the investigation of *Hypothesis 1* suggest that expert scorers seem to utilize a more holistic strategy for scoring that uses a less iterative decision making pattern than those used by non-experts. Based on the models presented in the literature review, experts seem to use a framework of scoring that is similar to the one proposed by Freedman and Calfee (1983). During the first phase of this framework of scoring, the scorer interprets the student writing through reading and reacting to the text. The purpose of this process is to create an image of the text, which is used during the second

phase of the scoring process. The second phase is dedicated to mapping the features of the student writing onto the scorer's interpretive framework of the scoring criteria by using a series of *review* processing actions. Through this process, judgments are made about how well the writer has demonstrated the various aspects of the scoring criteria, and a decision is formulated about the score to assign to the essay. In the third phase of the decision-making process, the scorer verbalizes this decision and cites evidence to justify that score.

On the other hand, intermediate and novice scorers seem to use a less fluent framework of scoring, one that is similar to the one described by Frederiksen (1992, April) for teacher evaluation. That is, non-experts seem to go through an iterative process of reading and *monitoring* portions of the essay. During each iteration, the scoreable features of that section of the essay are mapped onto the scorer's framework of writing. After completing this process for the entire essay, non-experts may *review* the essay prior to assigning a score. However, they are less likely to do so than are experts.

This interpretation of the results certainly does not suggest that the approach taken by non-experts is inferior to that taken by experts. It is likely that similar processing is occurring during the reading phase of an expert's scoring. However, more emphasis seems to be placed on the reading and comprehension process by experts than by intermediates and novices (Huot, 1993). It may be that experts have simply automated these procedures (Chi & Glaser, 1988). It is also possible that experts have adapted their scoring strategies to better accommodate a psychometrically-oriented scoring system (i.e., a system that focuses on maintaining high levels of quantitative indices of interrater reliability). Because scorers are often encouraged to work as accurately and quickly as possible in such a system, they may intentionally avoid using scoring methods that invite uncertainty or slow their pace. It may

be that such characteristics would not be observed in expert scorers who work in an *hermeneutic* scoring system--those that focus on presenting holistic and integrative evaluations of student work by preserving the idiosyncratic interpretations of individual raters (Moss, 1994).

Conclusion 2: There is little evidence to suggest that the manner in which frameworks of writing are employed by essay scorers is related to scoring proficiency. Scorers in this study demonstrated similar emphases in their content foci as has been observed in other studies of scoring (Diederich, French & Carlton, 1961; Freedman, 1979; Huot, 1993; Vaughan, 1991). That is, primary attention has been given to *storytelling, organization,* and *style*. Unfortunately, the analyses associated with *Hypothesis 2* provide little evidence that content focus has any relationship with scoring proficiency. An appealing explanation for the observed differences in processing action use by the proficiency groups is that these differences are caused by structural differences in the knowledge upon which those actions operate. Given the myriad of studies of expertise in other domains of human performance have indicated that the primary difference between experts and non-experts lies in the manner in which domain knowledge is structured (Berliner, 1986; Chase, 1983; Chi, Feltovich & Glaser, 1981; Chi & Glaser, 1988), a possible explanation for the failure to detect differences in the content focus of the proficiency groups in this study may be that there are methodological problems with the coding system. It may be that the coding system that was used is simply not sensitive to the kinds of differences in knowledge structures that make expert-like processing possible. Future studies should aim to determine whether there are differences in these knowledge structures that can be detected by other means of analysis.

Future Directions

Findings from this study suggest a number of avenues for further investigation. First, an important issue that has been left unresolved by this study is whether or not the frameworks of writing upon which scoring decisions are based differentiate scorers who demonstrate different levels of scoring proficiency. Future studies should aim to determine how the scoring criteria of expert and non-expert scorers are structured and whether this structure is related to the use of expert-like processing actions such as the ones observed in this study. This study has only indicated that the emphasis that is placed on different features of an essay during scoring is not related to the proficiency of the scorer.

Another research focus to which the findings of this investigation may be applied concerns studies of scorer recruitment and training. This study has shown that scoring proficiency in a *psychometric* scoring system (Moss, 1994) is highly evident in the manner in which an essay's contents are processed during evaluation. Previous efforts to train scorers have focused on developing the frameworks of writing of potential scorers. Little, if any, attention has been directed toward developing frameworks of scoring. Future training studies should aim to determine whether non-expert scorers can be trained to use expert-like approaches to scoring essays and whether the adoption of these strategies leads to improved scoring accuracy. Even if these training studies fail to improve scoring performance, it may be possible to use the findings of future studies to make quicker evaluative decisions about which scorer candidates are more likely to perform well on a scoring project.

References

Berliner, D.C. (1986). In pursuit of the expert pedagogue. *Educational Researcher, 15*(7), 5-13.

Chase, W.G. (1983). Spatial representations of taxi drivers. In D.R. Rogers & J.H. Slobada (Eds.), *Acquisition of symbolic skill* (pp. 391-405). New York, NY: Plenum.

Chi, M.T.H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chi, M.T.H. & Glaser, R. (1988). Overview. In M.H.T. Chi, R. Glaser, & M.J. Farr (Eds.), *The nature of expertise* (pp. xv-xxviii). Hillsdale, NJ: Lawrence Erlbaum.

Diederich, P., French, J.W. & Carlton, S. (1961). *Factors in judgments of writing ability.* (RB61-15). Princeton, NJ: Educational Testing Service.

Frederiksen, J.R. (1992, April). *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.R., Sipusic, M., Gamoran, M. & Wolfe, E.W. (1992). *Video portfolio assessment: A study for the National Board for Professional Teaching Standards.* Oakland, CA: Cognitive Science Research Center, Educational Testing Service.

Freedman, S.W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71(3)*, 328-338.

Freedman, S.W. & Calfee, R.C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S.A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75-98). New York, NY: Longman.

Huot, B.A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206-236). Cresskill, NJ: Hampton Press.

Kohr, R.L. & Games, P.A. (1977). Testing complex a priori contrasts on means from independent samples. *Journal of Educational Statistics, 2(3)*, 207-216.

Linn, R.L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15(1)*, 1-16.

Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Pula, J.J. & Huot, B.A. (1993). A model of background influences on holistic raters. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton Press.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex.

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.

Wolfe, E.W. (1995). *A study of expertise in essay scoring.* Unpublished doctoral dissertation, University of California, Berkeley, CA.

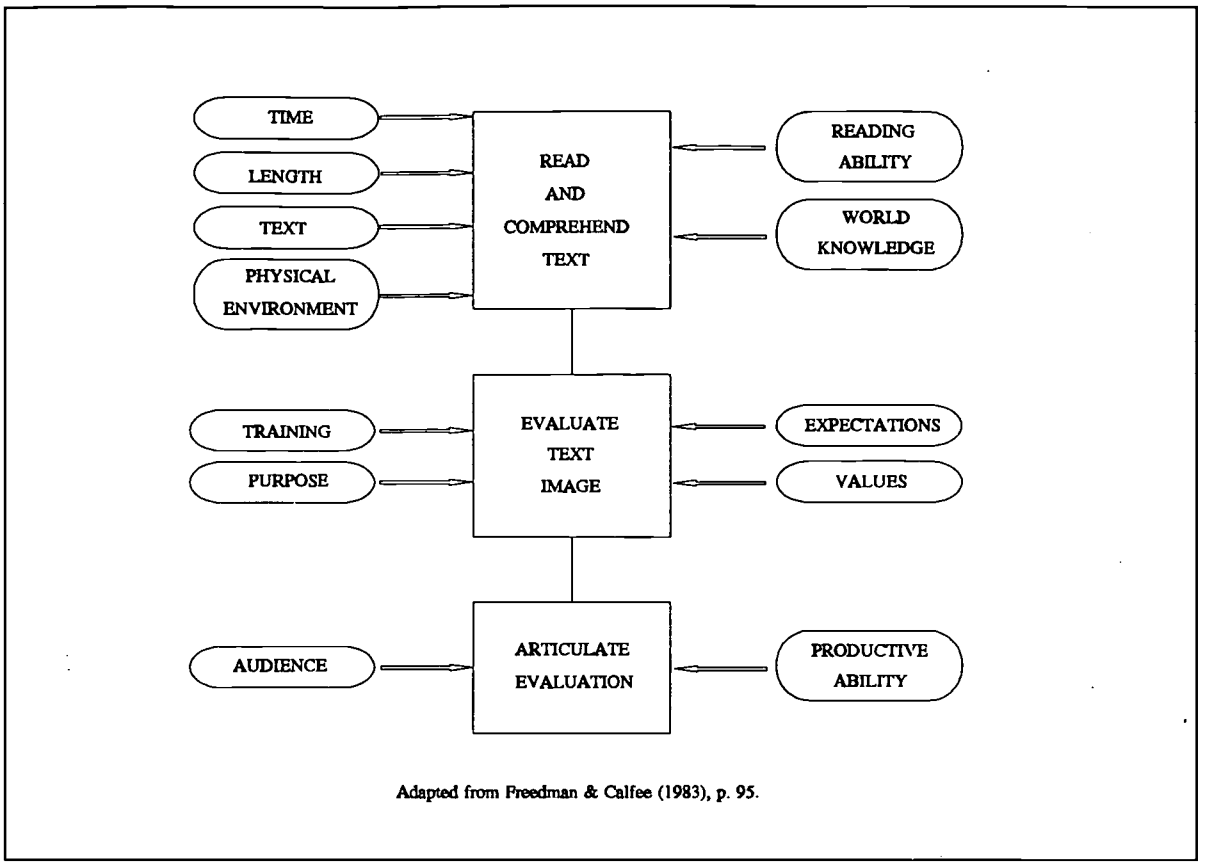

Figure 1: Freedman and Calfee Information-Processing Model of Scorer Cognition

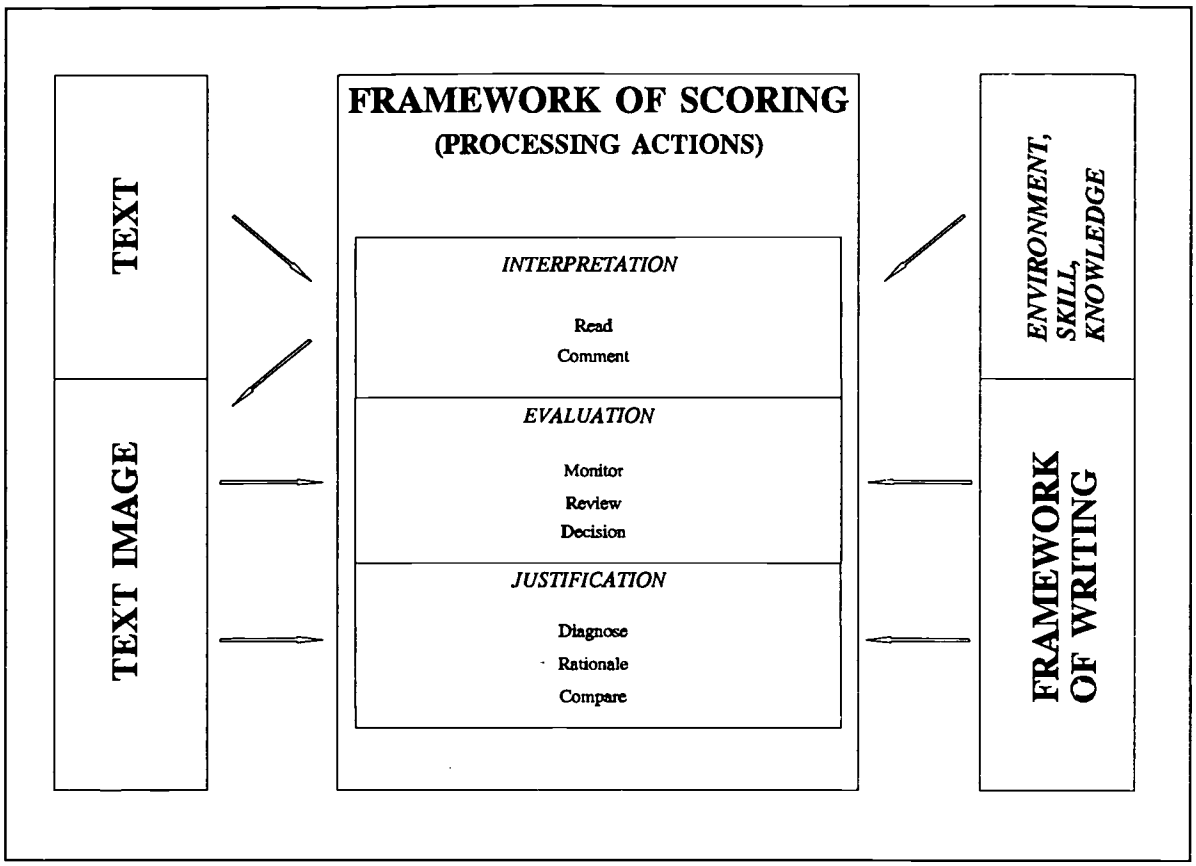

Figure 2: Expanded Model of Scorer Cognition

Table 1: Content Focus Categories for Essay Scoring

Content Category	Features	Definition
Appearance		Descriptions of the visual appearance of the text including descriptions of the legibility or length of the essay.
	Legibility	Descriptions of the readability of the text including references to the writing skills or handwriting quality.
	Length	Descriptions of the amount of writing contained in an essay including references to the number of pages.
Assignment		Descriptions of the extent to which the student addressed the writing assignment including references to the prompt, task or instructions.
Mechanics		Descriptions of the correctness of the writing at the word and phrase levels including references to spelling, punctuation, grammar or word usage.
Non-Specific		Descriptions of general or non-specific characteristics contained in the essay including the writer's control or skill.
	Control	Descriptions of the writer's control or command of the writing process.
	Skill	Descriptions of the writer that reference the general skills of the author.
Organization		Descriptions of the form of an essay including references to the focus, sentences or overall structure.
	Focus	Descriptions of the essay that reference the focus, flow, cohesion or direction of the writing.
	Form	Descriptions of the essay that refer to the use of organizational schemes (such as indentation, paragraphing or sections) in the writing.
	Structure	Descriptions of the overall organization or structure of the writing.

Table 1 (Continued): Content Focus Categories for Essay Scoring

Content Category	Features	Definition
Story-Telling		Descriptions of the characteristics that contribute to narration including references to communication or development of ideas, use of writing mechanisms or the story being told.
	Communication	Descriptions of the way the writer communicates ideas to the reader including references to the purpose or goal of the writer, whether the story is interesting, engaging or confusing, and the employment of ideas and their levels of sophistication or intelligence.
	Development	Descriptions of the way the writer develops ideas in the essay including references to the use of details for elaboration, the specificity of language or the support given for ideas expressed.
	Elements	Descriptions of the ways narrative and general writing mechanisms are used including references to action, scene or characters, or the use of dialogue or other language mechanisms.
	Story	Descriptions of the extent to which the narrative relays a story to the reader.
Style		Descriptions of the way a writer's individual style is relayed including references to the vocabulary used or evidence for a distinct writer's voice in the essay.
	Sentences	Descriptions of the complexity, control or structuring of sentences in the essay.
	Vocabulary	Descriptions of the use of vocabulary including references to word choice or wording.
	Voice	Descriptions of the distinct writer's voice including references to the writer's voice or style, the expression of the writer's emotions or the sophistication of thought contained in the writing.

Table 2: Processing Actions for Essay Scoring

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

Table 3: Group Proportions on Think Aloud Processing Actions

	Expert		*Intermediate*		*Novice*	
Processing Action	M_{prop}	SD_{prop}	M_{prop}	SD_{prop}	M_{prop}	SD_{prop}
Monitor	.06	0.06	.31	0.18	.24	0.25
Review	.57	0.18	.34	0.14	.33	0.27
Rationale	.27	0.13	.28	0.12	.33	0.25
Diagnose	.10	0.07	.07	0.04	.10	0.02

Table 4: Group Comparisons on Proportions for Think Aloud Processing Actions

Contrast	*v (v')*	*t*	*p*	*Method*
		Monitor		
$\mu_{experts}$ - 1/2($\mu_{intermediates}$ + $\mu_{novices}$)	(21)	-4.45	< .001	Welch's *t'*
$\mu_{intermediates}$ - $\mu_{novices}$	33	0.97	.34	*t*
		Review		
$\mu_{experts}$ - 1/2($\mu_{intermediates}$ + $\mu_{novices}$)	33	3.23	.005	*t*
$\mu_{intermediates}$ - $\mu_{novices}$	(18)	0.05	.996	Welch's *t'*
		Rationale		
$\mu_{experts}$ - 1/2($\mu_{intermediates}$ + $\mu_{novices}$)	33	-0.47	.64	*t*
$\mu_{intermediates}$ - $\mu_{novices}$	(16)	-0.62	.54	Welch's *t'*
		Diagnose		
$\mu_{experts}$ - 1/2($\mu_{intermediates}$ + $\mu_{novices}$)	33	0.52	.61	*t*
$\mu_{intermediates}$ - $\mu_{novices}$	33	-1.17	.25	*t*

Table 5: Group Proportions on Think Aloud Content Focus

	Expert		*Intermediate*		*Novice*	
Content Focus	M_{prop}	SD_{prop}	M_{prop}	SD_{prop}	M_{prop}	SD_{prop}
Appearance	.04	0.03	.06	0.06	.05	0.03
Assignment	.03	0.03	.02	0.02	.02	0.02
Mechanics	.09	0.04	.06	0.05	.10	0.04
Non-Specific	.16	0.08	.16	0.08	.13	0.05
Organization	.19	0.08	.24	0.10	.16	0.04
Storytelling	.34	0.05	.32	0.11	.41	0.07
Style	.15	0.05	.15	0.07	.14	0.04

Table 6: Group Comparisons on Proportions for Think Aloud Content Focus

Contrast	*v (v')*	*t*	*p*	*Method*
	Appearance			
$\mu_{experts}$ - $1/2(\mu_{intermediates} + \mu_{novices})$	33	-0.52	.61	*t*
$\mu_{intermediates}$ - $\mu_{novices}$	33	0.76	.45	*t*
	Assignment			
$\mu_{experts}$ - $1/2(\mu_{intermediates} + \mu_{novices})$	33	1.04	.31	*t*
$\mu_{intermediates}$ - $\mu_{novices}$	33	-0.46	.65	*t*
	Mechanics			
$\mu_{experts}$ - $1/2(\mu_{intermediates} + \mu_{novices})$	33	0.77	.45	*t*
$\mu_{intermediates}$ - $\mu_{novices}$	33	-2.21	.04	*t*
	Non-Specific			
$\mu_{experts}$ - $1/2(\mu_{intermediates} + \mu_{novices})$	33	0.49	.63	*t*
$\mu_{intermediates}$ - $\mu_{novices}$	33	1.22	.23	*t*
	Organization			
$\mu_{experts}$ - $1/2(\mu_{intermediates} + \mu_{novices})$	33	-0.43	.67	*t*
$\mu_{intermediates}$ - $\mu_{novices}$	(16)	2.60	.01	Welch's *t'*
	Storytelling			
$\mu_{experts}$ - $1/2(\mu_{intermediates} + \mu_{novices})$	(28)	-0.95	.35	Welch's *t'*
$\mu_{intermediates}$ - $\mu_{novices}$	33	-2.74	.01	*t*
	Style			
$\mu_{experts}$ - $1/2(\mu_{intermediates} + \mu_{novices})$	33	0.49	.53	*t*
$\mu_{intermediates}$ - $\mu_{novices}$	33	0.24	.81	*t*

TM 025513

AERA April 8-12, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Expert/Novice Differences in the Focus and Procedures Used by Essay Scorers	
Author(s): Edward W. Wolfe & Chi-wen Kao	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here
Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Edward W Wolfe	Position: Program Associate
Printed Name: Edward W. Wolfe	Organization: ACT
Address: ACT - 41 POB 168 Iowa City, IA 52240	Telephone Number: (319) 337-1548
	Date: 4/8/96

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 27, 1996

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a written copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (23)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 1996/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://tikkun.ed.asu.edu/aera/). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

---

[1]If you are an AERA chair or discussant, please save this form for future use.