DOCUMENT RESUME

ED 368 777 TM 021 208

AUTHOR	Wolfe, Edward W.; Feltovich, Brian
TITLE	Learning To Rate Essays: A Study of Scorer Cognition.
PUB DATE	Apr 94
NOTE	38p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Cognitive Processes; Criteria; *Essays; Evaluation Methods; *Evaluators; Experience; Holistic Approach; *Scoring; Standardized Tests; Writing Evaluation; *Writing Tests
IDENTIFIERS	Experts; *Mental Models; Performance Based Evaluation

ABSTRACT

This paper presents a model of scored cognition that incorporates two types of mental models: models of performance (i.e., the criteria for judging performance) and models of scoring (i.e., the procedural scripts for scoring an essay). In Study 1, six novice and five experienced scorers wrote definitions of three levels of a 6-point holistic scoring rubric at two times during a scoring project for a large-scale standardized writing assessment. Given practice and experience (3 days) in scoring, models of performance used by novices began to approximate those used by experienced scores. In Study 2, three better experienced scorers and three poorer experienced scorers engaged in a think-aloud task on three essays. This study revealed that better scorers stop more often while reading essays to comment and that the reading styles of better scorers are more consistent as a group. Better scorers are also more likely to make nonevaluative comments about the test itself or the writer, and they are more consistent in their use of content categories when discussing essay characteristics and more likely to focus on complex and abstract qualities (e.g., writer's voice or content development) than were poorer scorers. Six tables, three fugures are included. (Contains 25 references.) (Author/SLD)

Learning to rate essays:

A study of scorer cognition

Edward W. Wolfe

Brian Feltovich

American College Testing

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Running Head: SCORER COGNITION

Abstract

This paper presents a model of scorer cognition that incorporates two types of mental models: models of performance (i.e., the criteria for judging performance) and models of scoring (i.e., the procedural scripts for scoring an essay). Two studies were designed to investigate how scorers differ in their use of models of performance and models of scoring. In Study 1, six novice and five experienced scorers wrote definitions of three levels of a six-point holistic scoring rubric at two times during a scoring project for a large-scale standardized writing assessment (after completing training and after three days of scoring). This study revealed that, given practice and experience scoring, the models of performance used by novices began to approximate those used by experienced scorers. In Study 2, three better experienced scorers and three poorer experienced scorers were engaged in a think aloud task on three essays. Protocols were coded for processing actions and content focus. The study revealed that better scorers stop more often while reading essays to comment than do poorer scorers and that the reading styles of better scorers are more consistent as a group. Better scorers were also more likely to make non-evaluative comments about the text itself or the writer. With respect to content focus, better scorers were more consistent in their use of content categories when discussing essay characteristics. They were also more likely to focus on complex and abstract qualities (e.g., writer's voice or content development) than were poorer scorers.

Table of Contents

List of Tables

List of Figures

Learning to rate essays:

A study of scorer cognition

Purpose

Recently, much attention has been directed toward reliability issues in performance assessments. Ideally, scores from performance assessments used for high-stakes comparison and selection decisions will have levels of reliability comparable to those achieved for large-scale multiple-choice tests. If reliability is defined as the ratio of variance due solely to individual differences and the total observed variance (represented as σ_t^2/σ_x^2 in classical test theory; where σ_t^2 is *true score variance* and σ_x^2 is *observed variance*), then measurement error can be referred to in terms of sources of *construct-irrelevant variance* (i.e., facets that contribute to the magnitude of observed variance that do not contribute to true score variance). Reliability may be influenced by any of a number of sources of construct-irrelevant variance. Traditional assessment instruments have focused on stability of scores *over time* (i.e., test-retest reliability) and stability of scores *across items* (i.e., alternate forms reliability and internal consistency) as two sources of construct-irrelevant variance.

Because of the complex response formats of performance assessments they must be scored by people who are trained to make evaluative judgments about student responses. Due to differences in the backgrounds and values of scorers, methods used to train raters and the conditions under which scoring is performed idiosyncracies of each rater become another source of construct-irrelevant variance that must be controlled in order to insure high levels of reliability. As a result, the reliability of open-ended items is assessed both *between raters* (i.e., interrater reliability) and *within individual raters* (i.e., intrarater reliability). The purpose of this study is to investigate scorer variables that could potentially increase construct-irrelevant variance associated with the raters who score direct writing assessments.

Problem

A major hurdle for large-scale direct writing assessments to clear is achieving reliability levels that are adequate to allow wide-range comparisons and selection decisions to be made (Huot, 1990). The most influential sources of construct-irrelevant variance associated with performance assessments are instability of student performance across both time and items, and instability due to scorers. According to findings of generalizability studies, most of the construct-irrelevant variance in performance assessment scores is due to instability of student performance across prompts (Linn, 1993). The variability in observed scores on performance assessments that can be attributed to variation due to scorers is reported to be nearly nil.

However, these findings may be misleading. First, generalizability studies are usually performed using small groups of homogeneous scorers (e.g., graduate students from the same school of education) (Gao, Shavelson & Baxter, 1993; and Gao, personal communication, May 12, 1993). It may not be feasible to obtain comparable groups of scorers for nationally-administered large-scale assessments. Raters for these projects may be teachers from across an entire state or from across the nation. It is unlikely that the backgrounds of such scorers would be as similar as those used in generalizability studies. As a result, one would expect to see a larger contribution of scorer effects to the observed variance in large-scale performance assessments.

Second, a similar criticism can be made in reference to the training of scorers for generalizability studies. For most large-scale performance assessments, the cost of paying raters and the demand for short turnaround times reduces the time available to train scorers. As a result, one goal of training for large-scale performance assessments is to maximize interrater reliability while minimizing training and scoring time. Again, it is doubtful that generalizability studies are performed under such restrictive conditions. This results in homogeneous scoring performance and small scorer variance components.

Third, rater scores are not compared to "true score" estimates in generalizability studies. As a result, estimates of interrater agreement may be artificially inflated. If raters base scoring decisions on superficial or concrete characteristics of responses, such as text length or handwriting quality, scores from different scorers are likely to show high levels of agreement although their agreement with true score estimates (like scores of expert raters) would be lower (McColly, 1970). Again, this may be a particular problem with large-scale scoring projects in which a large number of scorers are trained in a short amount of time.

Unfortunately, little attention has been directed toward studying the influence of scorer effects on measurement error. Furthermore, the focus of psychometric studies of this phenomenon has been narrow, and these studies have been performed under rather idealistic conditions. We suggest that in order to understand how rater characteristics and reading styles contribute to construct-irrelevant variance in scoring performance assessments, it is necessary to identify differences in how raters score and how these differences relate to reliability estimates. With this focus in mind, we review the literature concerning scoring cognition in direct writing assessment and discuss a cognitive model upon which our two studies are based.

Theoretical Rationale

In an early study of the influence of raters on scoring, Freedman (1979) explored the focus that essay evaluators use when they rate papers. She manipulated several essays written by college students on four parameters: content, organization, sentence structure, and mechanics. Essays were created that contained combinations of strong and weak characteristics in each of these categories. Twelve evaluators rated 96 essays for strength in each category. An Analysis of Variance revealed that higher qualities of content and organization were associated with higher scores. Mechanics and sentence structure had smaller positive effects, and both showed a significant interaction with organization. An interesting secondary finding of this study was that training influenced the focus of scoring, suggesting that scorer prior experience may be one source of construct-irrelevant variance in direct writing assessment.

Vaughan (1991) performed the first direct study of the thinking of essay scorers. She asked nine experienced holistic raters to score six essays and to comment on them. Scorers were trained to use a six-point scale that emphasized organization, language expression, development, support, vocabulary, and mechanics. Five reading styles were identified: 1) *single-focus approach*, 2) *"first impression dominates" approach*, 3) *"two-category" strategy*, 4) *the laughing rater*, and 5) *a grammar-oriented rater*. Vaughan (1991) concluded that "the data show that raters are not *tabula rasa*. and do not, like computers, internalize a predetermined grid that they apply uniformly to every essay. Despite their similar training, different raters focus on different essay elements and perhaps have individual approaches to reading essays" (p. 120).

Huot (1993) explored the validity of holistic scoring as a reading process. He posited that raters' judgments emerge from their personal responses to the text and that holistic scoring may impede the fluency of this reading process. In his investigation, four expert raters and four novices read 42 essays aloud. Novices had never been trained or taken part in an holistic method of scoring. Experts all had prior holistic scoring experience and were trained specifically for this project.

With respect to rating essays, the two groups based their judgments on the same criteria: *content* and *organization* with some attention to *style* and *text appearance*. Expert scorers responded to the essay in a personal way more often than novices. They were also able to create meaning beyond their roles as evaluators. Although novices commented more times as a group, their remarks were limited to expressing their expectations, opinions, and judgments of the essays. Experts, on the other hand, were able to see themselves in the role of teacher or mentor. As for the reading process, experts used a more fluent reading process and tended to make their evaluations after--rather than while--reading. Novices interrupted their reading more often to make comments. Thus, the use of a scoring rubric may make it easier for raters to read and judge writing quality because it cuts down on the amount of cognitive activity necessary to make judgments about essay quality.

For the group of novices, there was little common agreement on the focus of rating. These raters also changed their strategies as they gained experience. Novices seemed to be searching for a set of criteria for scoring. Some of the novice strategies came from the essay itself ıather than from general knowledge about writing. Huot (1993) concludes that "what's important to remember in all of this is that the novice rater group did succeed, since novice raters used pretty much the same rating criteria as did their expert counterparts. The difference is that the novice effort to provide a good evaluation cost them any personal interaction with the student texts; all of their effort was diverted in assessing writing quality. ıp. 224)

Experts seemed to be satisfied with their individual rating style. Their styles revolved around a method of rating more than a criteria for judging writing quality. They used guidelines as a basis for their decisions, and appeared to have well-formulated strategies for applying the criteria and about the methods they should use. Expert styles included: 1) *conversing with the paper*, 2) *guarding against biases*, 3) *read rapidly to create an impression*, and 4) *being fair while focusing on content rather than mechanics*.

> It appears that the expert rater's stronger sense of method allowed him/her to reserve comments and to interact more personally with a student's writing than did the rating style of the novice rater group. An important difference in the models of rating for these two groups comes from the fact that for the most part the novice group had to improvise a strategy or rely on what it could do best. The expert group not only had the assistance that a scoring guideline provided, it also appears to have organized its past experience into a coherent set of rating strategies which helped to facilitate the personal involvement lost to those novice raters who had to improvise a rating strategy. (Huot, 1993, p. 226)

Huot concludes that there is no evidence that holistic scoring practices impede the ability of raters to read and assess the quality of student writing.

Pula & Huot (1993) examined the influence of background, training, and experience on holistic scoring. They interviewed 4 expert raters and 4 novices after they had scored 21 papers using an holistic scoring system. *Background reading* and *personal writing experience* seemed to be the most influential variables in determining the emphasis of scoring criteria. *Professıonal experience* was second only to *personal background* in its potential influence on rater focus. *Teaching* helped scorers understand how difficult some aspects of the writing process were to master. *Holistic scoring experience* helped the scorer to establish an internal rubric. The experience of reading thousands of papers, rather than exposure to a rubric, was cited as the most important aspect of scoring experience. *Professional training* was also cited as an important influence, along with the *primacy* (first), *recency* (latest), *novelty* (originality), and *repetition* (constancy) of the information provided in this training.

> (*The findings of this study*) suggest characteristics of people who might make gooa holistic scorers. Such individuals would have done the extensive reading necessary to give them an internal sense of the qualities of good writing, a real rubric. This study also suggests that teaching experience helps raters make an assessment of teachability, and that placement rating is best carried out by those who teach the courses into which they are placing students. Also important for individuals attempting to achieve scoring reliability with a group might be previous experience with holistic scoring or a commitment to the scoring task, with a concomitant willingness to agree with other members of the group on a negotiated external rubric. (Pula & Huot, 1993, p. 261)

The research cited here is supported by studies of experts in other fields, and it provides us with insights about the nature of the cognitive task of scoring essays. It is likely that scoring expertise develops over years of experiences as a result of exposure to a large number of examples of student writing. This allows the scorer to build a large repertoire of cases upon which generalizable principles may be realized by the expert (Gentner, 1988 and Pula & Huot, 1993). As a result, experts construct and organize their understandings of the domain in ways that are different from novices. More specifically, the schemata of experts are highly specialized, procedural and abstract. The concepts used by experts are connected to the conditions and procedures of their use (Glaser, 1987). In writing assessment, this may be evidenced by an expert having broader understandings of the role of an evaluator and the meaning of the writing outside of the assessment context (Huɔt, 1993). Furthermore, as a result of their varied experiences in the field, experts are more likely to be consistent in their interpretations of information, are more likely to realize overriding principles of the domain that guide them in decision making, and are more likely to recognize atypical patterns of information (Chi, Feltovich & Glaser 1981; Glaser, 1985; Huot, 1993; and Vaughan, 1991). The structure of the knowledge of an expert reduces the cognitive processing load, thus allowing them to automate thinking processes and utilize self-regulatory processes to control their thought and behavior patterns (Glaser, 1987 and Huot, 1993).

Unfortunately, these studies of scorer cognition have not been guided by a framework that can be used to evaluate the validity of generalizations like those cited above. It seems that an important next step for studies of scorer thinking is to generate a theoretical framework against which conflicting hypotheses can be compared so that our understanding of scorer cognition can move beyond a set of generalized statements of the characteristics of experts and novices to a deeper understanding of the conditions that influence scorer thinking and the relationships between various aspects of rater thinking.

In the remainder of this section, we formulate a model of scorer cognition. Ultimately, we hope to use this model to investigate ways of decreasing the amount of construct-irrelevant variance added to observed variance from scorer effects. In other words, we hope to use this model as a guide for exploring ways of reducing the contributions of scorer variables to the unreliability of direct writing assessments.

Interpretive Frameworks

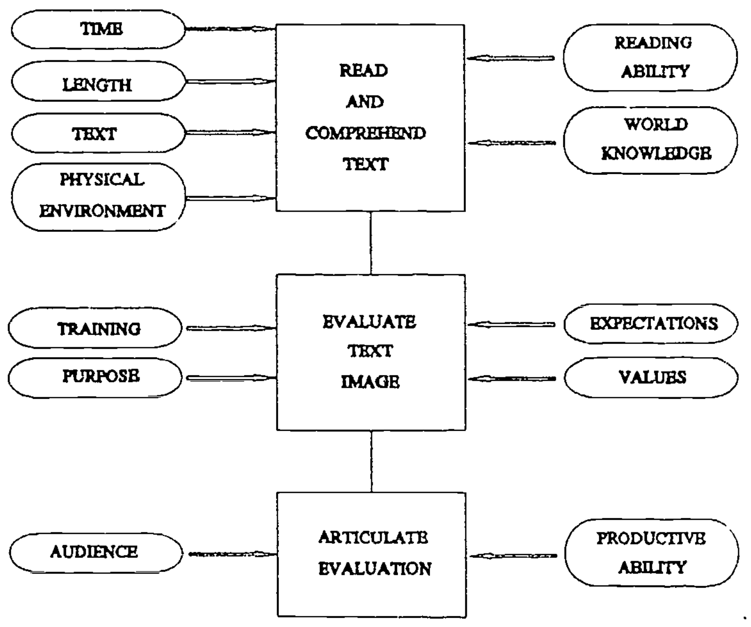
An *interpretive framework* is a cognitive representational structure. Most cognitive theories maintain that an individual constructs cognitive representations of objects and processes in order to understand the world. These representations serve as a point of reference for interpreting stimuli from the environment and for making predictions for the result of future courses of action. Typically, interpretive frameworks develop from naive and unsophisticated representations of reality to complex and holistic understandings (Chi, Glaser & Farr, 1988 and McDaniel & Lawrence, 1990).

In recent work in performance assessment, Frederiksen (1992) suggested that teacher evaluators use interpretive frameworks to understand and evaluate teacher performance. This notion of an interpretive framework suggests that the evaluator monitors a performance for some set of criteria. When a noteworthy instance of a performance criteria is observed, the evaluator makes a mental note of the aspect of the criteria being demonstrated and the degree of competence shown at that instance (*valence*). 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 assigns 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 the performance to others.

Voss and Post (1988) describe a similar use of an interpretive framework used by judges as they hear legal cases. During the *representation* phase of making a ruling, information relevant to the case is identified, the relationship between these pieces of information is defined, and each piece of information is weighted according to its importance for the decision at hand. After defining the realm of admissible evidence, the decision is made according to the rules established while representing the case. This stage is referred to as the *solution* phase. Finally, judges go through a *justification* phase in which a rationale is provided for the decision.

A similar process may be used by essay scorers as they judge writing quality. Freedman and Calfee (1983) describe an information processing model of essay scoring that is similar to the models described above. They identify three processes that are essential to rating 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 personal characteristics of the rater (e.g., reading ability, world knowledge, expectations, values, and productive ability) and environment characteristics (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 (Freedman & Calfee, 1983) shows how these processes interact on a global level.

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



Adapted from Freedman & Calfee (1983), p. 95.

In this model, information is taken from the printed text and an image of the response is constructed. This process is a "meaning making" activity during which the rater *interprets* student writing based on 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 rater's text image may be very different than one constructed by another rater. Based on the text image, the scorer compares 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 the writing sample. Finally, the evaluative decision is articulated through written or oral comments about the text.

We suggest that the complex decision-making activity of scoring essays, in which a number of interpretive frameworks are called upon to supply information to a number of cognitive processes, can be depicted by two primary components. These components, models of performance and models of scoring, define the interpretive frameworks that seem most relevant to the task of scoring essays.

Model of Perfcrmance

A *model of performance* is a cognitive representation of what constitutes proficient or non-proficient performance. In the case of direct writing assessment, it is a model of the characteristics indicative of writing ability and achievement (e.g., mastery of mechanics, ability to write with complex sentence structure, etc.). 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 model of performance are scoring rubrics, exemplar libraries, or other scorers. One would expect the model of performance constructed by a proficient scorer to be similar to the adopted scoring rubric, and that of a non-proficient scorer to be more dissimilar.

As suggested previously, the model of performance used by a particular scorer may not be adequately described by the scoring guide. For example, handwriting quality has long been suspected to influence holistic writing scores even when scorers are instructed not to base decisions on text appearance. A number of studies have shown handwriting neatness to be highly correlated with scores on papers even when papers of identical content were compared (Huck & Bounds, 1972; Markham, 1976; and McColly, 1970). Another textual feature that may be considered by scorers is essay length. Although it may be the case that the correlation of word counts and holistic scores is an artifact of the relationship of each of these variables with a mediating factor (Breland & Jones, 1984), Vaughan (1991) found that some raters mention this aspect of papers when scoring.

In our pilot study, we engaged two scorers in a think aloud task on a set of eight narrative essays. The raters were instructed to read each paper aloud and to verbalize any thoughts they had while scoring each paper. Typically, raters read the paper verbatim, occasionally interjecting comments like "'Their' is misspelled" or "I like that metaphor." After reading the paper, the scorer announced what score the paper should receive and provided a rationale for that decision. Many of the statements made by scorers directly referred to characteristics described by the scoring rubric.

These comments could be placed in one of the following categories (each corresponding to an element of the scoring rubric): 1) *Development* (ability to tell a story, add details and supporting ideas, and utilize language mechanisms), 2) *Organization* (ability to logically order a sequence of events), 3) *Voice* (ability to convey insight or personal style in a story, through the use of sentence structuring and words), 4) *Mechanics* (ability to control spelling, punctuation, capitalization, and language usage).

Some essay characteristics that the scorers cited did not fall into these categories. As a result, three new categories were created that accounted for the remaining comments. These categories are: 1) *Appearance* (the textual appearance of the essay), 2) *Subject* (compliance with the prompt or the effect a chosen topic has on compliance with the prompt) and 3) *Non-Specific* (general comments about the writing).

Taken together, these seven categories define a space of variables that essay raters might include in their models of performance for narrative writing. This classification system (referred to in Appendix A as *Content*) was adopted in the two studies described later in this paper to identify the characteristics of an essay that are most salient to individual scorers. In other words, we used the pilot study to create a coding system for identifying how a scorer's model of performance compares to the adopted scoring rubric and to the models used by other scorers.

Model of Scoring

The second interpretive framework relevant to essay scoring, a *model of scoring*, is a cognitive representation of the process through which one identifies and interprets evidence from a response and derives a score based on this information. It is a model of the manner in which a scorer manipulates a variety of knowledge representations during the decision-making process. In essence, a model of scoring serves as a script that the rater uses in order to insure fairness and accuracy in grading.

The model of scoring must contain a number of elements. First, it must represent the *process of interpretation*. How does one take in the stimuli and determine what aspects of the response will be considered as evidence for competence or non-competence? Second, the model of scoring must contain an understanding of the *process of evaluation*. It contains information concerning how discrepancies or inconsistencies in the evidence will be dealt with or how different aspects of the response will be weighted. A third component of a model of scoring is an understanding of the *process of documentation*. Because of the complexity of the scoring task, it is necessary to devise a system for recording scorer comments, scores, and for circulating responses among scorers. Finally, the model 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 our model of scoring, and these processes roughly correspond to those identified by Freedman and Calfee (1983). The process of *documentation* is not represented by their model but seems worthy of inclusion because of its importance in large-scale scoring projects. Neither is a fifth process alluded to in prior studies (Huot, 1993) which relates to personal involvement in the reading process (called *interaction*).

These five activities for rating essays were used by both scorers in our pilot study, and each of them seemed to be performed through the execution of a number of processes. As a result of combining these processes, the manner in which each scorer executed the script specified by the model of scoring appeared to be different. Therefore, in order to differentiate these general processes from the specific actions through which the scripts are enacted, the term *processing action* is introduced here.

Processing Action

A *processing action* is one of several cognitive activities that a scorer may perform when making a scoring judgment. Processing actions are used in combination to complete the script specified by the model of scoring. In the pilot study, not only were differences in content focus (e.g., development, organization, mechanics, etc.) identified, but each rater utilized these knowledge representations in different ways. For example, one rater tended to read the text from beginning to end, seldom stopping to note problems with it. He typically reviewed the text afterwards, citing content that subsequently factored into his scoring decision. The other scorer stopped often while reading to identify strengths and weaknesses in the essay and would occasionally state how these elements were helping him formulate an on-line decision prior to finishing the entire essay.

These observations lead to the identification of a number of cognitive tasks that raters perform during the decision-making process. During the *interpretive* phase of the model of scoring raters simply *read* the text to create a text image (occasionally summarizing or paraphrasing what they read). During the *evaluative* phase of the model of scoring, scorers often make a *decision* (assign a score or score range), *monitor* (note how the text image maps onto the model of performance), or *review* (survey how the text image maps onto the model of performance). For the *justification* phase, raters often *compare* elements of the text image to other sources of information, *diagnose* ways the essay could be improved, or provide a *rationale* to describe how the text image exemplifies certain aspects of the model of performance. The *interactive* process manifests itself through personal *comments* made by the scorer. These comments may refer to either the scoring process itself or to the text and the writer. *Documentation* processes were not observed because of the activity presented to subjects in the pilot study, but they might include processes like *record* a score, *change* a score, or *organize* scoring materials. The role of these actions in the coding system adopted for this study is described in the *Processing Actions* section of Appendix A.

The study of processing actions may provide insight into individual differences between scorers as they read an essay and formulate a scoring decision. Understanding these differences and how they relate to error variance in performance assessments could prove beneficial in identifying ways to control construct-irrelevant variance, thus increasing reliability of scoring. In the following studies, we identify differences in the ways scorers execute their models of scoring.

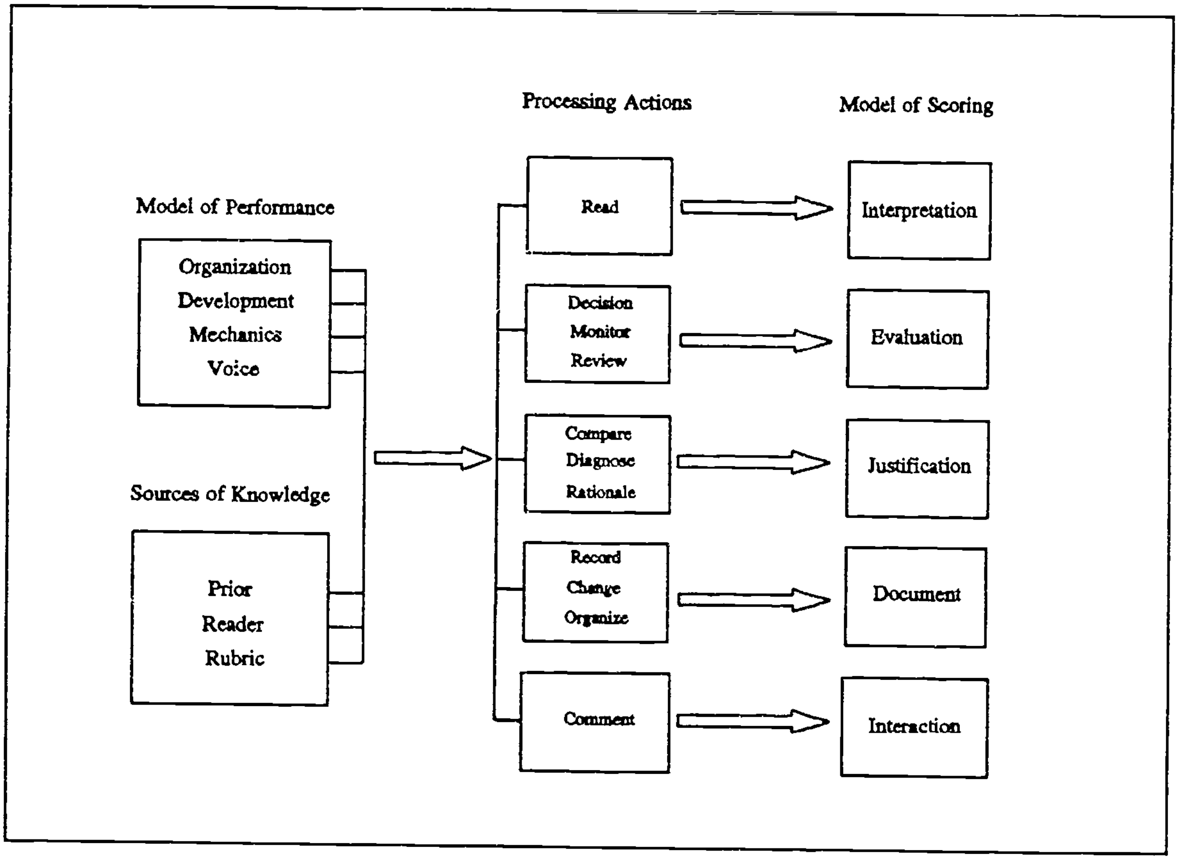
A Model of Scorer Cognition

To summarize, scoring cognition has been described here as an interplay of two primary components: knowledge representations and processing actions. Knowledge representations are classified as being either models of performance or models of scoring. A model of performance represents the criteria upon which scoring decisions are based. A model of scoring represents the process through which a scoring decision is made. Processing actions, on the other hand, are the tasks that are used to enact the model of scoring. They are executed according to the script contained in the model of scoring by manipulating various knowledge sources available to the scorer. Figure 2 shows how these components work together in the scoring process. It shows that the model of performance and other sources of knowledge feed information to the processing actions, which are executed according to the script specified by the model of scoring. Through this process, an image of the written text is created (interpretation), and its components are mapped onto the model of performance (evaluation). The evaluative decision is then monitored for fairness and accuracy (justification) and documented. Throughout the process, the essay scorer brings prior experience and knowledge to bear on the scoring activity in order to draw meaning from the text, scoring guide, and scoring process.

There is considerable evidence that the primary focus of models of performance in essay scoring is on the content and organization of an essay (Freedman, 1979; Huot, 1993) whether scorers have been given formal training or not (Pula & Huot, 1993). However, some scorers base their decisions on criteria other than that found in the scoring rubric (Vaughan, 1991), and individuals seem to emphasize different aspects of the content components (Huot, 1993). Furthermore, the breadth of possible considerations may be greater for less-experienced scorers than for experts (Huot, 1993 and Vaughan, 1991).

With respect to differences in the use of models of scoring, less information is available. The only findings here suggest 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). One would expect experienced raters to have more stable and efficient strategies for scoring that would allow them to become more personally involved in the reading activity (Huot, 1993). As a result, expert scorers are likely to read essays at a faster rate than novices (Pula & Huot, 1993).

Figure 2: Expanded Model of Scorer Cognition



Although 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 raters' models of performance, models of scoring, and use of processing options; it seems that 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. We conducted two studies of scorer cognition: one focusing on individual differences in the cognitive models used by essay scorers with different levels of experience and one focusing on how scorers with different levels of proficiency (accuracy) score. Both of these studies use model of scorer cognition described previously as a means to organize the results from our studies.

Study 1

In the first study, we examined how scorers' models of performance compared as a function of holistic scoring experience as well as how they change over a short period of time. Based on the literature on scoring cognition, one would expect experienced scorers to be more consistent as a group in their use of content concepts and that their conceptions would be more stable over time. Finally, one would expect the conceptions of performance held by novices to approximate those held by experienced scorers as a result of time and scoring experience.

Methodology 1

Eleven essay scorers (5 experienced and 6 novices) were trained to use an holistic scoring system for narrative essays prior to scoring a large number of student responses from a national administration of a standardized essay examination for eighth-grade students. All essays were written on the same narrative prompt. All scorers were asked to define the characteristics of papers found at levels One, Three, and Five of the six-point holistic scale. The scorers wrote these definitions both at the end of the end of two days of training and at the end of the third day of scoring. Written responses were transcribed to typing. Each complete thought written by a scorer was designated as a t-unit. Each t-unit was subsequently coded by the first author according to the *Content* section of the coding system described in Appendix A.

Results 1

Table 1 shows the frequency distribution of t-units across three dimensions: *content* (e.g., development, organization, etc.), *time* (i.e., during training or during scoring), and *level* of the scoring rubric (i.e., One, Three or Five). The relative use of content categories shows *development, organization* and *voice* being used the most and with *non-specific, mechanics, subject* and *appearance* being cited less often.

These analyses also afford an opportunity to observe how each content category was used to describe levels of the scoring guide. A Chi-square test revealed that content categories were used differently, depending on the level they were used to describe (χ^2 = 100.73, df = 12, p = 0.00). The catego. of *subject* and *appearance* were used more often to describe Level 1 papers than for the other levels. For Level 3, *organization* was cited in discussions of paper characteristics. *Development* and *voice* were discussed more often at Level 5 than for the other levels.

Table 1: Frequency of T-Unit Citation

		Content						
Time	*Level*	*Development*	*Organization*	*Voice*	*Mechanics*	*Subject*	*Appearance*	*Non-Specific*
Train	1	32	12	6	6	16	10	10
	3	53	31	12	5	8	2	10
	5	68	16	31	8	5	0	10
Score	1	24	5	3	3	18	7	11
	3	38	24	15	10	5	3	10
	5	48	10	20	8	3	3	11
Total		263	98	87	40	55	25	62

Group Comparisons

Few differences between experienced and non-experienced scorers were observed. Table 2 shows the overall frequency with which each group cited content categories for each level of the scoring rubric. There were only small differences in the use of specific categories by each group for each level (Level 1: $\chi^2 = 8.50$, df = 6, p = 0.21; Level 3: $\chi^2 = 8.47$, df = 6, p = 0.21; Level 5: $\chi^2 = 10.66$, df = 6, p = 0.10). However, the size of the p values for these statistics implies that there is some chance that these distributions do not come from a common population. Closer examination reveals that the experienced group used *development* more often in their descriptions of Level 3 and *voice* more often in their descriptions of Level 5. On the other hand, novice scorers used *organization* and *subject* more often in their descriptions of Level 1 and *voice* more often at Level 3. Finally, it should be noted that there were no differences in the use of the categories between the groups after training or during scoring (Training: $\chi^2 = 6.45$, df = 6, p = 0.38; Scoring: $\chi^2 = 2.81$, df = 6, p = 0.83). Although there were no significant differences between the groups at either time, after a few days of scoring the novice group's distribution was more similar to that of the experienced scorers (as indicated by the increase in the magnitude of the p value associated with the χ^2 statistic).

Table 2: Group Frequencies of T-Unit Citations

Group	*Level*	**Content**						
		Development	*Organization*	*Voice*	*Mechanics*	*Subject*	*Appearance*	*Non-Specific*
Experienced	1	30	4	4	3	20	10	11
	3	52	29	7	8	7	3	10
	5	58	11	36	6	3	1	11
Novice	1	26	13	5	6	14	7	10
	3	39	26	20	7	6	2	10
	5	58	15	15	10	5	2	10
	Total	263	98	87	40	55	25	62

Study 2

In the previous study, we examined how experienced scorers differed from novice scorers with respect to the characteristics of their models of performance. Although some small differences were observed, these findings do not help us identify why some scorers might outperform others on traditional measures of reliability. To begin with, experience does not necessarily equate with proficiency. Also, the validity of the retrospective report methodology is suspect (Nisbett & Wilson, 1977) because scorers are not likely to accurately report the knowledge used during post-scoring interviews. Finally, only content was examined in the previous study. It is likely that conceptions of content only play a partial role in determining scoring proficiency. In light of these problems, we performed a second study to identify how experienced scorers with varying levels of proficiency compared using a think-aloud methodology to elicit more realistic citations of content and execution of processing actions.

Methodology 2

Six professional scorers were individually engaged in a think aloud task on three narrative essays randomly-selected from a national administration of a large-scale standardized writing assessment. All raters were trained to use a scoring rubric for narrative essays, and each had scored a large number of essays holistically on at least one prior occasion. Raters were instructed to read each paper aloud and to verbalize any thoughts they had while arriving at a score. Protocols were tape recorded and later transcribed for analysis. Transcribed protocols were coded by the two authors using the coding system described in Appendix A. The protocols were first divided into t-units, and then each t-unit was assigned a process action rating and other associated information dictated by the coding scheme.

For the sake of comparison, scorers were divided into two groups: 1) *Better scorers* (those whose scoring performance was better than average on a large number of essays; that is, better than 60% perfect agreement with a randomly-selected second scorer) and 2) *Poorer scorers* (those whose scoring performance was lower than average; that is, lower than 60% perfect agreement). Table 3 shows the average scoring statistics for raters in each group. Although there was no difference in the number of papers read per hour by these groups, the better group averaged almost 9% more papers in perfect agreement with a randomly-selected second rater than the poorer group. Furthermore, only 2.9% of the papers scored by the better group were more than one score point away from the score assigned by a randomly-selected second scorer while the same statistic for the poorer group was 5.3%.

Table 3: Mean Rater Statistics for Proficiency Groups

	Statistic		
Group	*% Perfect*	*% Resolved*	*Papers read per hour*
Better	61.4	2.9	23.5
Poorer	52.4	5.3	23.4

Results 2

General Scorer Characteristics

In general, scorers began by reading the essay aloud, occasionally stopping to comment on its content. These comments took the form of either *comment* or *monitor* statements. Following the reading, the raters typically took one of two approaches to assigning a score: 1) *review* the paper to identify strong or weak aspects of the essay prior to assigning a score, or 2) assign a score and then provide a *rationale* for that score by citing strengths and weaknesses of the paper.

With respect to the use of processing actions, *monitor* was used most often. *Review* and *rationale* were used frequently but a little less often than was *monitor*. Scorers seldom used the *compare* and *diagnose* processing actions. With respect to *content* citations, the findings of Study 1 were replicated. *Development*, *organization*, and *voice* were cited most often as considerations for scoring. This is important from a validity standpoint because these three aspects are given the most emphasis in the scoring guide. *Non-specific* comments and *mechanics* were cited less often. Finally, *subject* and *appearance*, neither of which appear in the scoring guide, were cited least often by scorers.

Models of Scoring

The use of processing actions was compared between the two proficiency groups. 'Table 4 shows the average number of times each processing action was used per paper by raters in each group. The distributions of actions for the two groups were significantly different ($\chi^2 = 58.08$, df = 5, p = 0.00). The largest difference between these groups concerns the number of times scorers from each group stopped while *reading* an essay. The better group of scorers stopped their reading almost three times more often than scorers from the weaker group. In fact, the better group of scorers tended to interrupt their reading process to make comments after about every 50 words while poorer scorers tended to average about 150 words before stopping to comment.

Table 4: Mean Frequency of Processing Actions per Paper

	Process Action						
Group	*Read*	*Decide*	*Monitor*	*Review*	*Compare*	*Diagnose*	*Rationale*
Better	6.78	2.39	4.67	0.33	0.44	0.58	2.08
Poorer	2.89	1.89	1.44	3.67	1.67	0.89	2.33
δ	3.89	0.50	3.23	3.34	1.23	0.31	0.25

This difference is indicative of the reading strategies used by each group. Better scorers tended to consistently *monitor* the student essays. On the other hand, poorer raters (who stopped less often) relied most heavily on *review* and *compare* processing actions. Furthermore, the consistency of action use by poorer scorers was less than for the better scorers. Sixty-two percent of the processing actions used by better scorers were executed as *monitoring* actions. For the poorer scorers, only 39% of their processing actions fell into their most frequently used category: *review*.

With respect to the use of *comments* about the essays that were not evaluative in nature, although the numbers of comments made by each group were comparable, a much greater proportion of the total number made by poorer raters concerned their approach to scoring (32%), while the comments of better raters focused primarily on the writer or their personal reactions to the writing (90%) (so that only 10% of their comments related to scoring method). This difference is statistically significant ($\chi^2 = 7.22$, df = 1, p = 0.01).

Models of Performance

There were also small differences between the groups with respect to the focus they adopted when scoring essays. Table 5 shows the frequency distribution of comments made by scorers in each group. The distributions for the two groups were significantly different here also ($\chi^2 = 13.31$, df = 6, p = 0.04). Again, better scorers tended to be more focused in their concentration. In fact, an average of 51% of the comments made by each scorer in this group fell into the category of highest frequency for that scorer. For the poorer raters, only 34% of the total number of comments made fell into the category used most often by each individual. As a group, better scorers tended to focus on *voice* and *development* in making decisions while poorer scorers tended to make more comments about *mechanics*.

Table 5: Mean Frequency of Content Citations per Paper

Group	**Content**						
	Development	*Organization*	*Voice*	*Mechanics*	*Subject*	*Appearance*	*Non-specific*
Better	3.33	1.67	1.78	0.33	0.22	0.00	0.56
Poorer	2.33	2.33	1.00	1.11	0.22	0.44	1.00
δ	1.00	0.66	0.78	0.78	0.00	0.44	0.44

Discussion

We have identified four general conclusions that are supported by the results of these two studies. *First, the results suggest that the thinking that scorers engage in while scoring essays is driven by their models of performance.* That is, scorers use specialized schemata to represent the criteria upon which scoring judgments are based. In both of our studies, the primary focus of these models of performance was on the development of ideas, the organization of content and the writer's voice. Mechanics and textual appearance received less attention. These findings are also supported by Freedman (1979), Vaughan (1991), and Huot (1993). Furthermore, our studies suggest that the models of performance adopted by our scorers portray writing as a developmental process in which a scorer's concerns focus on rather simple writing characteristics at earlier stages (e.g., compliance with the prompt and textual appearance) to focusing on more complex characteristics at later stages (e.g., the emergence of content development and a writer's voice). This depiction of writing as a developmental process reflects the characteristics emphasized in the scoring rubric used by our scorers.

Second, a scorer's model of performance becomes more cohesive and complex as the result of exposure to student writing and practice in scoring. Pula and Huot (1993) suggest that the criteria used by experienced scorers is more consistent because of their exposure to a wide variety of issues, concepts and practice in the field of writing. We found that better scorers are more likely to apply their models of performance consistently across student responses and over time than were novices and poorer scorers. This may mean that the criteria used by better scorers has been "internalized" over their repeated exposure to examples of student writing (Gentner, 1988). And, like experts in other fields (Chi, Feltovich & Glaser, 1981), our poorer scorers were more likely to focus on concrete and simplistic characteristics of student papers (e.g., textual appearance) while better scorers are more likely to focus on characteristics that are more complex and abstract (e.g., emergence of a writer's voice). However, we also found that, given experience and practice, the models of performance used by novices are likely to approximate those used by experts.

Third, essay scorers tend to use a model of scoring similar to those identified in other fields that require expert judgements. That is, the process used to score essays resembles the process used by experts in other fields who engage in judgment-making. Voss and Post (1988) suggest that the decision-making of judges relies on three general processes: interpretation of the information presented, evaluation of the evidence, and justification of the decision. Our scorers demonstrated the same general process when scoring student essays. They interpreted the content of the text by reading the essay. They either monitored the essay's content or reviewed it in order to evaluate its quality. And, the scorers justified their decisions by providing a rationale for the assigned score. They also interacted with the text at various times throughout this process as suggested by Huot (1993) by making non-evaluative comments about the scoring process or about the text and the writer.

Fourth, the model of scoring used by essay scorers may differentiate better scorers from poorer ones. Although Huot (1993) found that experienced scorers are more likely to read an essay from beginning to end, seldom stopping to comment on its content, we found that this strategy was more characteristic of poorer experienced scorers. Our better scorers tended to monitor the content of the essay during the interpretive phase of the decision-making process. This may be consistent with other studies of expertise which suggest that experts spend more time in the initial stages of problem representation (Glaser & Chi, 1988). But, it may not be consistent with the notion that an expert's thinking is more automated than that of a novice (Glaser, 1987). Because this finding is not easily interpreted, further study is warranted.

Our study did reveal other, more easily interpreted findings concerning the processes that might differentiate expert scorers from other scorers. We found that better scorers were more consistent as a group in their use of processing actions. That is, their models of scoring were more similar, as a group, than those of poorer experienced scorers. We also found that better scorers were able to extend their thinking beyond the scoring task by becoming more personally engaged in the text. This finding is supported by the work of Huot (1993).

Our two studies suggest that the model of scorer cognition presented in this paper is an adequate representation of both the processes and the knowledge that are called upon by professional essay scorers. We also conclude that developmental differences, similar to those demonstrated by experts and novices in other fields, may be characteristic of essay scorers. The most prominent differences shown in our studies concern the models of performance (criteria) and models of scoring (procedural scripts) used by scorers. Expert scorers may have more cohesive and complex models of performance, and they may use more consistent and automated models of scoring.

The next step in the development of this model should focus on identifying other developmental differences so that the model can be defined more specifically and expanded to include characteristics of developmental stages in the acquisition of scoring expertise. Hopefully, with added knowledge and understanding of the nature of the scoring process, we can create better scoring and training procedures so that the measurement error associated with open-ended assessment items can be greatly reduced.

References

Breland, H.M. & Jones, R.J. (1984). Perceptions of writing skills. *Written Communication, 1(1),* 101-119.

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. & Farr, M.J. (Eds.). (1988). *The nature of expertise.* Hillsdale, NJ: Lawrence Erlbaum.

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.R., Sipusic, M., Gamoran, M. & Wolfe, E.W. (1992). *Video portfolio assessment: A study for the National Board for Professional Teaching Standards.* Emeryville, 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). Longman: New York, NY.

Gao, X. (May 12, 1993) personal communication.

Gao, X., Shavelson, R.J. & Baxter, G.P. (1993). *Generalizability of a state-wide performance assessment.* Paper presented at the Annual Meeting of American Educational Research Association, Atlanta, GA.

Gentner, D.R. (1988). Expertise in typewriting. In M.H.T. Chi, R. Glaser & M.J. Farr (Eds), *The nature of expertise* (pp. 1-33). Hillsdale, NJ: Lawrence Erlbaum.

Glaser, R. (1985). *Thoughts on expertise* (Technical Report No. 8). Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh.

Glaser, R. (1987). Thoughts on expertise. In C. Schooler & K.W. Schaie (Eds.), *Cognitive functioning and social life structure over the life course* (pp. 81-94). Norwood, NJ: Ablex.

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

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.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communications, 41(2),* 201-213.

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.

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

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

McColly, W. (1970). What does educational research say about the judging of writing ability? *The Journal of Educational Research, 64(4),* 148-156.

McDaniel. E. & Lawrence, C. (1990). *Levels of Cognitive Complexity: An approach to the measurement of thinking.* New York, NY: Springer-Verlag.

Nisbett, R.E. & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

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.

Voss, J.F. & Post, T.A. (1988). On the solving of ill-structured problems. In M.T.H. Chi, R. Glaser & M.J. Farr (Eds.), *The nature of expertise* (pp. 261-285). Hillsdale, NJ: Lawrence Erlbaum.

Appendix A: Coding System for Scorer Think Aloud Protocols

The Model of Scorer Cognition

The model of scorer cognition described in this paper is a conceptual mapping/information-processing model of an essay scorer's decision making process. In order to document the rater's portrayal of this model, a think aloud activity is used with essay raters as they score a number of essays. It is assumed that the statements made 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

From a scorer's protocol, each complete and non-repetitive though is designated as a thought-unit (t-unit). Each t-unit can be coded with respect to a number of dimensions. For example, a statement indicates 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* classification or *source* of knowledge). 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. Also the rater may provide some a comment that is of non-evaluative *type i*. 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 scorer can be coded according to the action being executed. An *action* refers to one of several processes that a scorer 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 process. 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). Each of these classes of actions requires a certain type of knowledge in order to be executed. Table 6 shows the classifications of actions and the specific actions associated with these classes as well as the types of knowledge that may be required by each action.

Table 6: 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 rater'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 rater'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 rater'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	*Type*

Content

Content plays an important role in the decision making of an essay scorer. *Content* refers to the language and values contained in the scorer's model of performance that is used as the "rules" for making scoring decisions. The scorer's model of performance is called upon to supply information when a rater 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 scorer's model of performance.

For our purposes, the following content sources 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, 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

Scorer 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) *Scorer* (scoring of the paper is compared to scores that might be assigned by other scorers) 3) *Rubric* (paper is compared to descriptions provided in the rubric).

Type

Interactive processing actions (i.e., *comments*) are performed by relaying information that is not specific to the rating process. A scorer may make a comment about the strategy used to arrive at a score. Scorers 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 *types* of scorer 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 rater 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 6.

The scorer reads 141 words from the essay.

The scorer 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 scorer reads the remaining 289 words in the essay.

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

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

Figure 3: 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		Source	Type	Valence
Interpret:	Read	Prior	Scoring	Positive (+)
Evaluate:	Decision (V), Monitor (C,V), Review (C,V)	Scorer	Reading	Neutral/Fail (N/F)
Justify:	Compare (S), Diagnose (C,V), Rationale (C,V)	Rubric		Negative (-)
Interact:	Comment (T)			Range (1-6)

Content

Appearance	Development	Mechanics
Non-Specific	Organization	Subject
Voice		

Rater ID: ME

Paper ID: 1234567