Cognitive Differences in Proficient and Nonproficient Essay Scorers

EDWARD W. WOLFE, CHI-WEN KAO and MICHAEL RANNEY

Written Communication 1998 15: 465
DOI: 10.1177/0741088398015004002

This article examines the behavioral differences of essay scorers who demonstrate different levels of proficiency for a psychometric scoring task. The authors compare three proficiency groups to identify differences in (a) essay features they consider, (b) their understandings of the scoring rubric, and (c) their decision-making procedures. Results indicate scorers with different levels of proficiency do not focus on different essay features when making evaluative decisions but their understandings of the scoring criteria may vary. Proficient scorers are more likely to focus on general features of an essay when making evaluative decisions and to adopt values espoused by the scoring rubric than are less proficient scorers. Also, proficient scorers make evaluations by reading the entire essay and then reviewing its content, whereas less proficient scorers may interrupt the reading process to monitor how well the essay satisfies the scoring criteria. Finally, the authors discuss implications for scorer selection and training.

Cognitive Differences in Proficient and Nonproficient Essay Scorers

EDWARD W. WOLFE
University of Florida
CHI-WEN KAO
Kentucky Department of Education
MICHAEL RANNEY
University of California, Berkeley

When essays are evaluated in large-scale assessment settings, scorers make judgments about how well specific pieces of student work

Authors' Note: Portions of this research were funded by ACT, Incorporated and a postdoctoral fellowship at Educational Testing Service. The authors thank John Frederiksen and Mark Wilson for their input on the analyses of these data and Isaac Bejar, Joan Heller, and Mark Reckase for their input on an early draft of this manuscript. Portions of this article were presented at the annual meeting of the American Educational Research Association in New York, New York (April 1996) and at the International Conference on the Learning Sciences in Evanston, Illinois (July 1996).

WRITTEN COMMUNICATION, Vol. 15 No. 4, October 1998 465-492

demonstrate writing competence. Such judgments often result in differences of opinion between two scorers. In a psychometric scoring system (i.e., one that emphasizes maintaining high levels of quantitative indicators of consistency among scorers), differences of opinion are seen as potential sources of measurement error and indicate a need for further training or refinement of the scoring rubric and training materials (Moss, 1994). These differences of opinion may indicate that scorers think differently about the features of the essay on which scores are based and about the procedures used to read and evaluate the essay. The purpose of this study is to determine whether scorers' thinking patterns are associated with proficiency for a large-scale, psychometric essay scoring task. The literature concerning essay scoring is not conclusive about why some scorers are better able to come to agreement than are others. However, it does suggest specific variables that might account for individual differences in scorer cognition. Drawing on the literature concerning essay scoring and expert-like decision making, we hypothesize that rater proficiency in a psychometric scoring task may manifest itself in the cognitive behaviors of scorers. After describing our investigation of these hypotheses, which employed a think-aloud methodology, we then present the results of our investigation. Finally, we suggest several ways in which the results of our study can be used to improve the identification and training of scorers for large-scale writing assessments.

THEORETICAL BACKGROUND

Scorer Cognition

In the literature concerning scoring, little attention has been directed toward scorer cognition. Much of the literature that focuses on scoring emphasizes the development and application of methods for training scorers (Charney, 1984; White, 1985), evaluating the contribution scorers make to measurement error (Brennan, 1992; Lane, Liu, Ankenmann, & Stone, 1996; Shavelson, Baxter, & Gao, 1993), or detecting and correcting scorer errors once they have been committed (Engelhard, 1994; Linacre, 1989; Lunz, Wright, & Linacre, 1990). Although each of these lines of research examines a different way to reduce scoring errors, we developed this study to extend knowledge about essay scoring by identifying ways to decrease the occurrence of

scorer errors through scorer selection and training. Thus, we began by looking at differences in the thoughts and behaviors of scorers.

In their research on essay scorer cognition, Pula and Huot (1993) examine how scorers' prior experiences influence the way scorers learn and subsequently apply a scoring rubric to essays. Their work describes how the foundation for a scorer's understanding of a scoring rubric is laid by identifying three experiential factors that differentiate expert from novice scorers. The first factor is personal background; a scorer's previous experiences as a reader and a writer create a frame of reference for future thinking about writing. The second factor that influences a scorer's capacity to learn and apply a scoring rubric is the scorer's professional training. For example, the primacy and recency of the training, the uniqueness of the perspectives presented by one's mentors, and the amount of repetition encountered during professional training may all influence a scorer's understanding of a particular scoring rubric. Third, work experience (e.g., experiences in teaching writing and scoring essays) seems to influence how well a scorer is able to adopt a predetermined scoring rubric.

Vaughan (1991) examines individual differences in a more direct way by employing a think-aloud task to study the thinking patterns of experienced essay scorers. From analyses of the protocols, Vaughan concludes that scorers do not internalize uniformly a predetermined scoring rubric. Despite similar training, different scorers may focus on different features of the essay or may focus their evaluations on essay features that are not cited in the scoring rubric. Vaughan also concludes that scorers have individualistic approaches to reading essays that may vary widely from one scorer to another. Huot (1993) extends Vaughan's (1991) work by comparing the evaluative methods and foci used by both experienced and nonexperienced essay scorers. Think-aloud protocol comments generated by essay scorers were coded according to their processing actions (i.e., the procedures used to make the scoring decisions, such as making personal comments or reviewing an essay's content) and scoring foci (i.e., the features of the essay on which scoring decisions are based, such as organization or grammar). Huot (1993) finds that although individual scorers may emphasize different aspects of an essay when making a scoring decision (as suggested by Vaughan, 1991), there do not seem to be systematic differences between the scoring foci of experienced and nonexperienced scorers. Huot also finds that experienced scorers are more consistent as a group in their emphases on essay features for individual essays. With respect to the use of processing actions, Huot shows

that novice scorers take less personal interest in essays than do experienced scorers. Personal responses to the texts by experienced scorers are more common and more varied than those of novices. Finally, novices tend to state expectations for the writing during their evaluations, especially in ways that interrupt the reading process. Experienced scorers, on the other hand, use more fluent reading processes and tend to make their evaluations after, rather than while, reading the essay.

The studies performed by Pula and Huot (1993), Vaughan (1991), and Huot (1993) may be useful in scorer selection and training contexts because they identify a variety of factors that may influence a scorer's capacity to adopt a scoring rubric and use it when reading and responding to student writing. However, two problems arise in making such applications: First, the studies all focus on how experience rather than proficiency relates to raters' behaviors. Second, none of the studies offer causal explanations of how and why certain thought patterns result in lower levels of agreement between raters.

With respect to the second of these problems, Frederiksen (1992) presents a model of scorer thinking that describes how teacher evaluators use interpretive frameworks (i.e., internalized representations of the qualities of a good teacher) to understand and evaluate teacher performance. In this model, the evaluator uses the interpretive framework as a filter for monitoring a teacher's classroom performance for criteria that are deemed important by the evaluator. When a noteworthy example of a performance criterion occurs, the evaluator makes a mental note of the characteristics demonstrated and the degree of teaching competence shown at that instance. Thus, the interpretive framework serves as a means for understanding and recognizing the parameters of the performance under assessment. Differences among evaluators' interpretive frameworks would therefore lead to different interpretations of a teacher's performance. After all noteworthy moments have been observed, the evaluator considers all of the observations, mentally assigns weights to them, and decides what score to assign. The last step in the process is to create a rationale. Thus, the interpretive framework also is used to organize and communicate ideas about teachers' performances to others.

Freedman and Calfee (1983) describe an alternative approach to scoring in their information-processing model of essay scoring. Their model identifies three processes that are essential to evaluating a composition: (a) reading text to build a text image, (b) evaluating the text image, and (c) articulating the evaluation. In their model, infor-

mation is taken from the printed text and a mental image of the student's writing (i.e., text image) is constructed. The scorer interprets the writing based on world knowledge, beliefs and values, and knowledge of the writing process. The conditions of the physical reading environment also may influence the form that the text image takes. As a result, the text image is not an exact replica of the original text, and one scorer's text image may be very different from the text images constructed by other scorers. Based on the text image, the scorer compares various aspects of the writing to internalized representations of the scoring criteria (cf., Frederiksen's [1992] notion of an interpretive framework). Through this process, judgments are made about the text and a decision is made about how well the writer has demonstrated writing competence. Finally, the evaluative decision is articulated through written or oral comments about the text.

The models described above suggest three possible causes of differences of opinion among scorers. First, Frederiksen's (1992) model suggests (as does the research of Pula and Huot, 1993) that different conclusions may be drawn by scorers who adopt different scoring foci (i.e., internalized representations of the scoring criteria). Second, the Freedman and Calfee (1983) model adds to this the notion that different scorers may create individualistic text images (i.e., interpretations of the content presented by the student) and may base their evaluations on these different variations. Third, as implied by the differences in the Freedman and Calfee (1983) and the Frederiksen (1992) models and as made explicit in the research of Vaughan (1991) and Huot (1993), scorers may use different processing actions to read and evaluate the essay. For example, the model described by Frederiksen (1992) portrays scoring as an iterative process in which a scorer makes multiple evaluative decisions, each of which revises the previous one. The Freedman and Calfee (1983) model, on the other hand, describes a more 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.

Operationalizing Scorer Cognition

Elsewhere, Wolfe (1997) proposes a model of scorer cognition that summarizes how these three components (text images, scoring foci, and processing actions) may relate to each other. In the current study, we explore how two of these components (scoring foci and processing actions) manifest themselves in the cognitive behaviors of scorers who

exhibit different levels of scoring proficiency. Our goal is to identify relationships that may be used to guide scorer selection and training decisions in large-scale, psychometric scoring projects. In this section, we operationalize several aspects of scoring foci and processing actions and formulate hypotheses about how the behaviors of scorers with different levels of proficiency in a psychometric scoring system may vary.

Ericsson and Simon (1993) describe a think-aloud methodology that has proven useful in tracking mental processing in a number of fields of human performance. Generally, this methodology assumes that the thought processes involved in the performance of cognitive tasks can be described as a series of mental states, each of which is the end product of the processing of information. When thinking occurs, the information and cognitive procedures that lead to a particular state may be brought to attention and reported verbally. These verbal reports are considered to be valid only to the extent that verbalization of the information does not interfere with the performance of the tasks. Fortunately, concurrent verbal reporting does not seem to interfere significantly with one's ability to perform tasks that require verbal reasoning (such as essay scoring). While thinking aloud, participants are asked to report information as it is heeded in attention (not to explain what they are doing). Based on a coding system that the researcher derives through an analysis of the various thought sequences and information that may be called into play while performing the task, the statements contained in the resulting think-aloud protocol are analyzed. In the following sections, we describe how a think-aloud methodology can be used to expose the scoring foci and processing actions used by essay scorers.

Scoring Focus

We use the term *scoring focus* to refer to a scorer's mental weighting of the various components of the scoring criteria. The scoring focus adopted by a scorer may be determined by (a) interactions between the scorer's prior beliefs and understandings of writing and the writing process, (b) the compatibility of the writer's values and the scoring rubric, and (c) the effectiveness of the methods and materials used to train the scorer (Frederiksen, Sipusic, Gamoran, & Wolfe, 1992; Pula & Huot, 1993). As a result, the scoring focus adopted by a particular scorer may differ from the explicit, external criteria contained in the scoring rubric. In a psychometric scoring system, one

would expect the scoring focus adopted by a highly proficient scorer to be very similar to the external scoring rubric because a scorer is trained to abandon previously held values and to adopt those espoused by the rubric writers. Recall that Huot (1993) found few differences between the scoring foci used by experienced and nonexperienced essay scorers. In a psychometric scoring system, initial small differences are likely to diminish even further because scorers undergo extensive training. Scorer training procedures in psychometric scoring projects indoctrinate scorers into a system in which they become like minded. Many of the training activities commonly used in large-scale assessment settings model the process of focusing on specific essay features during discussions of prescored examples of student work (i.e., benchmarks). In such environments, it seems unlikely that even minimally proficient scorers would dramatically diverge from the scoring rubric in the weighting of scoring-focus categories when making scoring decisions. Hence, we investigated the following family of hypotheses (i.e., hypotheses that are related in terms of their content and intended use; Kirk, 1995) concerning the relationship between scorer proficiency and scoring focus.

Family 1: Scorers who demonstrate different levels of scoring proficiency do not weight essay features differently when making evaluative decisions.

Elsewhere, Wolfe and Feltovich (1994) identify four categories to portray the essay features that may be weighted differentially by a narrative essay scorer: (a) mechanics, (b) organization, (c) storytelling, and (d) style. Each of these categories may be evidenced by a scorer during a think-aloud task. Table 1 contains a detailed description of how these scoring-focus categories are defined in this article. Although these categories are strongly influenced by the scoring rubric adopted for this study, the categories identified by Wolfe and Feltovich (1994) are similar to those identified by other researchers using different scoring systems and scoring rubrics (Barritt, Stock, & Clark, 1986; Breland, 1983; Breland & Jones, 1984; Diederich, French, & Carlton, 1961). A sample protocol, coded according to these scoring-focus categories, is shown in the first column of the table in Appendix A.

Although Family 1 proposes that scorers in a psychometric scoring system will exhibit few differences in their use of scoring-focus categories, it seems unlikely that scorers who demonstrate different levels of proficiency with a particular scoring rubric will have identical

Table 1
Scoring-Focus Categories

Scoring-Focus	Definition
Mechanics	Descriptions of the correctness of the writing at the word and phrase level (including references to spelling, punctuation, grammar, and word usage)
Organization	Descriptions of the form of an essay, such as the focus (including flow, cohesion, and direction of the writing), paragraphs (including references to the use of organizational schemes such as indentation, paragraphing, and sections), and overall structure (including general organization)
Storytelling	Descriptions of the characteristics that contribute to narration, such as the communication of ideas (including references to the purpose and goals of the writer; whether the story is interesting, engaging, or confusing; and the level of sophistication of the ideas presented), development of ideas (including references to the use of details for elaboration, the specificity of language, and the support given for the ideas expressed), use of writing mechanisms (including references to action, scene, or characters and the use of dialogue and other language mechanisms), and the extent to which the writer tells a story
Style	Descriptions of the way a writer's individual style is relayed through the use of sentence structure (including the complexity, control, and structuring of sentences), the sophistication of the vocabulary used (including references to word choice and wording), and evidence for a distinct writer's voice in the essay (including references to the writer's voice and style, the expression of the writer's emotions, and the sophistication of thought contained in the writing)

representations of that rubric. As we describe below, scorers' mental representations of the scoring rubric may differ in ways other than how the components of the scoring rubric are weighted.

Much of the research concerning expert-like decision making in domains other than essay scoring suggests that one of the most striking differences between experts and novices lies in the structure of their domain-relevant knowledge. Numerous researchers have concluded that expert knowledge is organized better and focuses on principles (rather than on facts) in the domain of performance. As a result, experts approach cognitive tasks by identifying the important principles on which the task is based. On the other hand, novices are more likely to focus on the surface details of the task (Glaser & Chi, 1988). In the domain of essay scoring, one would expect the knowl-

Table 2
Specificity Categories

Specificity	Definition
General	Descriptions that reference elements of the text or text image in broad terms without identifying specific, concrete examples of those features
Specific	Descriptions that identify specific and concrete examples of elements contained in the text or text image

edge structures used by proficient essay scorers to resemble those used by experts in other domains. Hence, one would expect essay scorers who are proficient in a psychometric scoring system to focus on general essay features when making evaluative decisions. Less proficient scorers, on the other hand, should be more likely to focus on specific essay features during their evaluations. We offer the following family of hypotheses in light of this notion.

Family 2: Scorers who demonstrate higher levels of proficiency in a psychometric scoring system are more likely to exhibit general (and less case-specific) understandings of the criteria on which essays are evaluated, whereas scorers with lower levels of proficiency are more likely to focus on specific essay features.

Based on preliminary work by Wolfe (1995), we use the term *degree of specificity* to refer to the extent to which a particular scorer's use of the scoring-focus categories demonstrates principle-driven versus surface-level applications of the scoring rubric. When asked to perform a think-aloud task, a scorer might make general references to the qualities of an essay that are considered when making an evaluative decision. For example, a scorer who uses a general reference to mechanical errors may say something like this when evaluating an essay: "This paper contains a lot of errors in spelling and punctuation." On the other hand, a scorer might make more specific references to essay features while scoring. Such a scorer would make comments about mechanics, such as "The student misspelled *their* and used a semicolon incorrectly in that sentence." Table 2 formally defines these categories for depicting the degree of specificity demonstrated by specific scorer comments. The sample protocol in Appendix A illustrates coding according to these degree of specificity categories.

Another way that the structure of scorers' mental representations of the scoring rubric may differ concerns the extent to which scorers adopt the rubric writers' language and values. Little research has focused on the extent to which essay scorers are able to set aside their personal values and adopt scoring rubrics generated by others. Pula and Huot (1993), however, suggest that this may be an important characteristic of successful essay scorers. In a psychometric scoring system, scorers are expected to maintain high levels of agreement both with other scorers and with scores assigned by test developers to validity papers (i.e., papers that have been prescored and are sent through the scoring system for monitoring purposes). In large-scale assessment practice, heavy emphasis tends to be placed on interrater agreement. Nevertheless, scorer training employs prescored examples of student writing, often accompanied by descriptions of how specific essay features map onto the scoring rubric. As a result, one would expect more proficient essay scorers to be more likely to adopt the language and values represented in the scoring rubric (i.e., a rubric-centered perspective). Less proficient scorers might be more likely to use the language and values that they bring with them to the scoring task (i.e., a self-generated perspective). Hence, we set forth the following family of hypotheses concerning the degree of rubric adoption.

Family 3: Scorers who demonstrate higher levels of proficiency in a psychometric scoring system are better able to internalize the language contained in the scoring rubric (i.e., use a rubric-centered perspective), whereas scorers with lower levels of proficiency are more likely to use a self-generated perspective.

We use the term *degree of rubric adoption* to refer to the extent to which an essay scorer employs the descriptive phrases contained in the scoring rubric when discussing an essay's quality. When engaged in a think-aloud task, a scorer typically uses key words or phrases to refer to features of the essay that he or she considers when arriving at a scoring decision. These key words and phrases may be either rubric centered or self-generated. Statements that are rubric centered are either quoted or paraphrased sections of the scoring rubric that a scorer is trained to use. Self-generated statements, on the other hand, focus on essay features that either are not specifically referenced in the scoring rubric (e.g., scorers often mention features such as length of text or handwriting quality although they are not mentioned in the scoring rubric) or are subclasses of the features contained in the rubric (e.g., scorers may mention errors in spelling, punctuation, and usage

Table 3
Rubric-Adoption Categories

Rubric Adoption	Definition
Self-generated	Descriptions that reference elements of the text or text image using terms and phrases that either are not contained in the rubric or depart from the intended meaning of those terms and phrases as designated by test developers
Rubric centered	Descriptions that reference elements of the text or text image using terms and phrases that are contained in the rubric and conform to the intended meaning of those terms and phrases as designated by test developers

although the rubric only discusses mechanical errors in general). Table 3 formally defines these categories for depicting rubric adoption. The sample protocol in Appendix A illustrates coding according to these degree-of-rubric-adoption categories. The scoring rubric used by scorers in this study is shown in Appendix B.

Processing Actions

Another prevailing conclusion drawn from the literature concerning expert-like decision making is that experts are able to perceive large patterns of information, allowing them to solve problems without performing exhaustive searches for solutions. As a result of this ability to chunk information, experts have better short-term memory, are able to automate routine processing tasks, and are able to perform cognitive tasks very quickly (Glaser & Chi, 1988). In the domain of essay scoring, Wolfe (1997) finds that proficient scorers in a psychometric scoring system are better able to withhold judgment when scoring essays, a finding supported by the work of Huot (1993) and Pula and Huot (1993). As is true for other domains that require expert judgment (Voss & Post, 1988), it is likely that such a trait would manifest itself in the ways that processing actions are used by essay scorers by allowing more proficient scorers to use interpretive and evaluative procedures that facilitate both text comprehension and holistic judgments. Less proficient scorers, on the other hand, may need to break the evaluative task down into more manageable tasks (i.e., iteratively reading and evaluating the text). Hence,

Family 4: Scorers who demonstrate higher levels of proficiency in a psychometric scoring system are better able to handle the cognitive demands of scoring by reading the texts to build text images and then evaluating the text images, whereas scorers with lower levels of proficiency will use an iterative procedure to break the texts down into pieces that are more easily managed.

We use the term *processing action* to refer to the specific procedures that essay scorers use to interpret texts, evaluate them in light of the scoring rubric, and assign and justify scores for them. The manner in which processing actions are used by scorers may be determined by background knowledge, prior scoring experiences, and reading skills (Freedman & Calfee, 1983). These experiences and skills lay the foundation for the scorer to develop a mental script of the process through which (a) text images are created and compared to the scoring criteria and (b) the accuracy of scores is evaluated. The structure of this script may be inferred from the specific procedural steps (i.e., processing actions) that scorers make when assigning scores to essays in think-aloud settings.

In their pilot study, Wolfe and Feltovich (1994) document a number of processing actions that seem to be central to essay scoring. For example, scorers often monitor how the texts or text images map onto the scoring rubric while they are reading the texts by making mental notes about specific features of the essays. Many scorers also seem to review the essays after they finish reading them (i.e., take stock of how the texts or text images map onto the scoring rubric). Scorers may also diagnose ways that the essays could be improved by citing how specific weaknesses could be corrected. Scorers may also provide a rationale for a particular decision by describing how the text images exemplify certain aspects of the scoring rubric. In the context of Family 4, one would expect proficient essay scorers to use review processing actions as they map features of the text images onto the scoring criteria. Less proficient scorers, on the other hand, would resort to monitoring as they break the evaluative task down into an iterative series of reading a portion of the text and mapping that section onto the scoring criteria. Table 4 contains a detailed description of how the processing-action categories are defined, and Appendix A provides a sample protocol that is coded using these categories.

Table 4
Processing-Action Categories

Processing Action	Definition
Monitor	Descriptions that reference elements of the text or text image in terms of a scoring-focus category during reading (i.e., making notes)
Review	Descriptions that reference elements of the text or text image in terms of a scoring focus category after completing the reading (i.e., taking stock)
Diagnose	Descriptions that reference the shortcomings of the text or how it could be improved in terms of a scoring-focus category
Rationale	Descriptions that reference elements of the text or text image in terms of scoring-focus category as support for an assigned score

METHOD

In the preceding section, we laid the foundation for a study of essay scorer cognition by describing several relevant cognitive characteristics that may differentiate scorers of different levels of proficiency within a psychometric scoring framework (i.e., one that emphasizes maintaining high levels of quantitative indicators of score reliability and validity). We established four families of hypotheses, each associated with a different variable that might explain the cognitive foundations of individual differences in scorer proficiency. Table 5 summarizes these variables and the associated families of hypotheses. The remainder of this section defines the methodology of the study we designed to test these hypotheses.

Scorers

Participants ($n = 36$) were selected from a pool of essay scorers ($N = 60$) who took part in a large essay-scoring project. All scorers were trained to use a 6-point holistic scoring rubric for narrative writing (Appendix B) in a manner common to many large-scale writing assessments (e.g., Fowles, 1978). Scorers read the scoring rubric, discussed examples of student writing that fell into each level of the rubric, and scored calibration sets to determine how well the scorers had mastered the rubric. After training, scorers began scoring a large number (about 6,500) of narrative essays written by 10th graders drawn to be a representative sample of U.S. public and private school students. Students responded in writing to a prompt that asked them to describe a time when they were scared. Each essay was scored by

Table 5
Families of Hypotheses

Family	Variable	Hypothesis
1	Scoring focus	Scorers who demonstrate different levels of scoring proficiency do not weight essay features differently when making evaluative decisions
2	Degree of specificity	Scorers who demonstrate higher levels of proficiency in a psychometric scoring system are more likely to exhibit general (and less case-specific) understandings of the criteria with which essays are evaluated, whereas scorers with lower levels of proficiency are more likely to focus on specific essay features
3	Degree of rubric adoption	Scorers who demonstrate higher levels of proficiency in a psychometric scoring system are better able to internalize the language contained in the scoring rubric (i.e., use a rubric-centered perspective), whereas scorers with lower levels of proficiency are more likely to use a self-generated perspective
4	Processing actions	Scorers who demonstrate higher levels of proficiency in a psychometric scoring system are better able to handle the cognitive demands of scoring by reading the text to build a text image and then by evaluating that text image, whereas scorers with lower levels of proficiency will use an iterative procedure to break the text down into pieces that are more easily managed

two scorers selected at random from the larger pool of scorers. At the end of the second day of scoring, an intraclass correlation (Shrout & Fleiss, 1979) was computed for each of the scorers in the larger pool. The intraclass correlation (r_{ic}) was computed to indicate the agreement between the scores assigned to all essays that were scored by individual scorers and the scores assigned to these same essays by the randomly selected second scorers (about 150 essays per scorer).

Three groups of participants (12 per group) were selected randomly from the distribution of interrater agreement indices. These groups represented the lower, middle, and upper thirds of the distribution of raters. Competent raters showed relatively low levels of agreement with other scorers, with an average $r_{ic} = .74$. Intermediate raters showed relatively middle levels of agreement with other scorers, with an average $r_{ic} = .80$. Proficient raters showed relatively high levels of agreement with other scorers, with an average $r_{ic} = .87$. To determine whether these proficiency groups were indeed statistically distinct from each other, the intraclass correlations were trans-

formed to the equivalent Fischer z value, and the average transformed values were compared via *t*-tests. These comparisons revealed that although the difference between the competent and intermediate scorers was not statistically significant, $t(22) = 1.64$, $p = .06$, the effect size was moderate ($r_{pb}^2 = .11$). We assume that the failure to achieve statistical significance using this sample size is simply a result of the small sample size and that the competent and intermediate scorers indeed are different from one another. The difference between the intermediate and proficient scorers, on the other hand, was both statistically significant and meaningfully large, $t(22) = 2.33$, $p = .01$, $r_{pb}^2 = .20$.

Procedures

From the 6,500 essays in the scoring project, 24 essays were randomly selected. Each of the 36 participants performed a think-aloud task while scoring these 24 essays during a private interview session. Interviews lasted approximately 1.5 hours each. Participants were given think-aloud instructions and practice consistent with the guidelines suggested by Ericsson and Simon (1993). Protocols were tape recorded, and coding was performed after the interview by two individuals who had prior experience both working with essay scorers and analyzing think-aloud protocols. Coders parsed each statement made by a scorer into complete and independent units. Each unit was coded according to four dimensions: (a) the essay feature referenced (i.e., scoring focus), (b) the degree of specificity of the statement, (c) the degree of rubric adoption demonstrated, and (d) the cognitive task performed (i.e., processing action). Appendix A illustrates a sample coded protocol. Each coder independently coded two thirds of the data; thus, one third of the data were coded by both coders. Cohen's kappa (κ) (Liebetrau, 1983) was computed for each coding dimension, and intercoder agreement was deemed acceptable ($\kappa = .93$ for scoring focus, $\kappa = .87$ for degree of specificity, $\kappa = .91$ for degree of rubric adoption, and $\kappa = .85$ for processing actions).

Analyses

To compensate for individual differences in verbosity, counts for each coding dimension were converted to proportions. That is, counts for individual coding categories (e.g., mechanics) for each coding

dimension (e.g., scoring focus) were summed across essays and divided by the total number of statements made by a participant for that coding dimension. These proportions served as the data for group comparisons. Two-sample *t*-tests were performed to investigate the four families of hypotheses (i.e., Family 1, Family 2, Family 3, and Family 4), with each family composed of several statistical hypotheses. Our goal was to identify monotonic relationships between the cognitive structures and activities used by scorers with different levels of proficiency within the psychometric scoring system (i.e., $\mu_{Proficient} > \mu_{Intermediate} > \mu_{Competent}$ or $\mu_{Proficient} < \mu_{Intermediate} < \mu_{Competent}$). To this end, a pair of a priori orthogonal contrasts were applied to the proportions for each coding category of scoring-focus, degree-of-specificity, degree-of-rubric-adoption, and processing-action coding dimensions. The two contrasts compared proficient scorers to intermediates (Equation 1) and intermediates to competents (Equation 2). Family-wise error rate was corrected using the sequentially rejective method described by Holm (1979). For statistically significant differences, we examined the proportion of variation accounted for by group differences. That is, we relied on the squared point-biserial correlation (r_{pb}^2) as an index of the importance of the size of an unlikely result. As noted by Cohen (1988), effect sizes of .25 are relatively large.

$$\Psi = \mu_{Proficient} - \mu_{Intermediate} \quad (1)$$

$$\Psi_2 = \mu_{Intermediate} - \mu_{Competent} \quad (2)$$

RESULTS

Our results support all four of our families of hypotheses. As expected, the most proficient scorers did not differ in the scoring foci that they applied to narrative essays (recall Family 1). However, the least proficient scorers, those in the competent group, tended to place a slightly heavier emphasis on storytelling than did those in the other two groups. With respect to the degree of specificity (recall Family 2), proficient and intermediate scorers made more comments about general characteristics of the essays, whereas competent scorers were more likely to mention specific features of the writing in their evaluations. Proficiency groups also differed in the degree of rubric adop-

Table 6
Descriptive Statistics of Group Proportions for Scoring-Focus Categories

Scoring Focus	Proficient	Intermediate	Competent
Mechanics	.12 (.05)	.08 (.07)	.13 (.05)
Organization	.23 (.09)	.32 (.13)	.20 (.06)
Storytelling	.44 (.09)	.41 (.11)	.50 (.07)
Style	.20 (.06)	.19 (.08)	.17 (.05)

NOTE: $n = 12$ for each group. Means are shown with standard deviations in parentheses. Two of the eight contrasts associated with this family of hypotheses were large enough to be considered important: organization (intermediate-competent) and storytelling (intermediate-competent).

tion they demonstrated in their discussions of essays (recall Family 3): our proficient scorers were more likely to use rubric-generated vocabulary in their discussions, whereas intermediate and competent scorers were more likely to rely on self-generated descriptions of essay features in their evaluations. Furthermore, proficiency groups used different processing actions as they made their evaluative decisions (recall Family 4). Proficient scorers tended to use a holistic approach to scoring essays, reading the whole essay and then basing their evaluations on their understanding of its entirety. Intermediate and competent scorers, on the other hand, were more likely to use a bottom-up approach during the evaluation, reading short sections of the essay and making evaluative decisions about each of those smaller units.

Scoring Focus

Family 1, which predicted that proficiency groups would not focus on different essay features when evaluating essays, was supported partially the data. Table 6 shows the means and standard deviations of the proportion of statements coded into each of the four scoring-focus categories for each proficiency group. The most commonly cited scoring feature for all three proficiency groups was storytelling, and the least common feature mentioned was mechanics. The means show only minor variations between groups across the four coding categories. *T*-tests revealed that only two of the eight contrasts associated with this family of hypotheses were large enough to be important. The difference between intermediate and competent scorers was statisti-

Table 7
Descriptive Statistics for Group Proportions for Degree-of-Specificity Categories

Degree of Specificity	Proficient	Intermediate	Competent
General references	.83 (.10)	.82 (.07)	.73 (.11)
Specific citations	.17 (.10)	.18 (.07)	.27 (.11)

NOTE: $n = 12$ for each group. Means are shown with standard deviations in parentheses. One of the two contrasts associated with this family of hypotheses was large enough to be considered important: general/specific (intermediate-competent).

cally significant, with a large effect size for both organization, $t(22) = 2.90$, $p = .008$, $r_{pb}^2 = .28$, and storytelling, $t(22) = 2.39$, $p = .03$, $r_{pb}^2 = .21$. These results suggest that although there are few differences between proficient and intermediate scorers in terms of the scoring foci that they adopt, there are slight differences between these two groups and the competent scoring group. The latter tends to place a heavier emphasis on storytelling in our scoring rubric, whereas the former two place a heavier emphasis on organization.

Degree of Specificity

Family 2 also was supported by our data. We predicted that more proficient scorers would demonstrate more generalized applications of the criteria on which essays were evaluated, whereas less proficient scorers would be more likely to focus on surface details of the essays. Table 7 shows the means and standard deviations of the proportion of statements coded as general references and specific citations for each proficiency group. The group means show that all three groups were more likely to make general references in their discussions of essay quality but that competent scorers were more likely to make specific citations than were the other two groups. The difference between intermediate and competent scorers is statistically significant, with a moderate effect size, $t(22) = 2.39$, $p = .03$, $r_{pb}^2 = .21$).

Degree of Rubric Adoption

Family 3, which predicted that scorers who are more proficient within a psychometric scoring task would be better able to internalize the scoring rubric than would less proficient scorers by making more

Table 8
Descriptive Statistics of Group Proportions for Degree-of-Rubric-Adoption Categories

Degree of Rubric Adoption	Proficient	Intermediate	Competent
Rubric centered	.47 (.10)	.34 (.13)	.34 (.12)
Self-generated	.53 (.10)	.66 (.13)	.66 (.12)

NOTE: $n = 12$ for each group. Means are shown with standard deviations in parentheses. One of the two contrasts associated with this family of hypotheses was large enough to be considered important: rubric/self (proficient-intermediate).

rubric-centered statements about essays during their evaluative discussions, was supported by our data as well. Table 8 shows the means and standard deviations of the proportion of statements that were coded as rubric centered and self-generated for each proficiency group. As shown by the group means, self-generated statements were predominant in all three groups. However, nearly half of the statements made by proficient scorers fell into each rubric-adoption category, whereas intermediate and competent scorers gave self-generated descriptors about two thirds of the time. This difference (proficient-intermediate) is both statistically significant and meaningfully large, $t(22) = 2.75$, $p = .01$, $r_{pb}^2 = .26$.

Processing Actions

We found strong support for Family 4, which predicted that scorers with higher levels of proficiency would demonstrate a better ability to handle the complex cognitive task of scoring essays by using more holistic scoring methods. That is, we expected proficient scorers to use a read-then-evaluate approach to scoring, whereas less proficient scorers would use an iterative read-evaluate-read-evaluate approach to scoring. As a result, we expected to observe differences in review and monitor processing-action use between the three proficiency groups. As shown in Table 9, our findings supported these predictions. The means of the proportion of statements that were codified into each processing-action category show only small between-group differences on the rationale and diagnose processing actions. However, very large differences between proficient scorers and intermediate scorers were observed for the review and monitor processing-action categories. For both of these categories, the proficient-intermediate

Table 9
Descriptive Statistics of Group Proportions for Processing-Action Categories

Processing Action	Proficient	Intermediate	Competent
Monitor	.06 (.06)	.31 (.18)	.24 (.25)
Review	.57 (.18)	.34 (.14)	.33 (.27)
Diagnose	.27 (.13)	.28 (.12)	.33 (.25)
Rationale	.10 (.07)	.07 (.04)	.10 (.06)

NOTE: $n = 12$ for each group. Means are shown with standard deviations in parentheses. Two of the eight contrasts associated with this family of hypotheses were large enough to be considered important: monitor (proficient-intermediate) and review (proficient-intermediate).

comparisons were statistically significant and meaningfully large: monitor, $t(22) = 4.56$, $p = .0002$, $r_{pb}^2 = .49$; review, $t(22) = 3.49$, $p = .002$, $r_{pb}^2 = .36$.

CONCLUSIONS

Our results suggest several things about the relationship between the cognitive characteristics of essay scorers and proficiency with a psychometric scoring task. First, our results suggest that essay scorers in a psychometric scoring system seem to focus on and place similar weights on various essay features when making evaluative decisions (recall Family 1). (However, as noted above, competent scorers showed a slight tendency to diverge from the two more proficient groups of scorers, placing more emphasis on the storytelling elements of the scoring rubric.) Second, although all scorers weigh essay features similarly, they do seem to consider these essay features at different grain sizes (recall Family 2). That is, scorers who demonstrate higher levels of proficiency for the psychometric scoring task cite features of essays that are more general, whereas scorers who are less able to agree with others focus on more specific features of the essay. Third, there seems to be a relationship between psychometric scoring proficiency and the extent to which a scorer adopts language that is consistent with the scoring rubric (recall Family 3). Proficient essay scorers seem more likely to employ the language used by test developers in their descriptions of essay quality. Less proficient scorers, on

the other hand, are more likely to use descriptive words that are not found in the scoring rubric. This suggests that proficiency in a psychometric scoring system may be related to one's ability to adopt the values espoused by the test developers. What is not clear from our results is whether our proficient scorers were actually better able to abandon their previously held values (i.e., those that they brought to the scoring project) or whether the proficient scorers in our study simply entered into training with personal values that were more similar to those held by the test developers. Fourth, there seems to be a relationship between psychometric scoring proficiency and the scoring procedures (i.e., processing actions) that scorers use (recall Family 4). More specifically, proficient psychometric essay scorers seem to use a top-down approach to scoring essays, through which they build overall images of the texts and make holistic judgments of the writing quality. On the other hand, less proficient scorers seem to use a bottom-up approach to scoring, breaking the decision-making task down into an iterative series of read-evaluate procedures.

These results have several implications for scorer selection, training, and monitoring in a psychometric scoring system. Certainly, further research is necessary to determine why some raters are better able than are others to adopt the language of the scoring rubric. Although our study has shown that proficient scorers are more likely to use the same language as test developers, it is unclear whether this fact can be used to help guide rater-selection decisions in large-scale scoring projects. Are raters who come to the scoring project with similar values and beliefs about writing better able to adopt the scoring rubric? Or, are some raters better able to abandon their previously held beliefs in lieu of those being presented during scorer training? Once this determination is made, efforts to train scorers could be facilitated by selecting only the scorers from a pool of potential candidates who show the most promise for adopting the scoring rubric with the least amount of training—a practice that could significantly reduce the cost of large-scale essay-scoring projects.

Our results also suggest that scorer training, as it is practiced currently in many large-scale settings, may be accomplishing its goal. Many large-scale scoring projects approach scorer training with the purpose of creating a group of like-minded individuals who look for and focus on similar essay features when making scoring decisions. Scoring trainers accomplish this goal by discussing prescored examples of student works and identifying features of essays that warrant the scores assigned by a committee of test developers. Our results

suggest that these goals seem to be realized because our scorers, particularly the most proficient, tended to weight essay features similarly when making scoring decisions. Two additional questions should be explored in future research concerning scorer training: How well can scorers be trained to alter the procedures that they use when making scoring decisions? Furthermore, does the adoption of scoring procedures that are more similar to those used by proficient essay scorers lead to more accurate scoring? To our knowledge, few scorer-training programs explicitly provide scorers with training concerning the procedural actions that one should use when making an evaluative decision (one exception is discussed by Wolfe, Gitomer, & Carter, 1998). If these two questions can be answered affirmatively, then it may be possible to increase scoring accuracy beyond the levels that are observed currently in large-scale essay scoring projects. However, until such research is performed, we can only speculate about the teachability of the procedural knowledge used by essay scorers. It is possible that the efficient processing demonstrated by proficient scorers is enabled only once certain declarative knowledge structures (e.g., content knowledge or knowledge concerning students' writing characteristics) are in place.

Regardless, we believe that our results suggest an interesting potential alternative method for monitoring, evaluating, and training scorers. In current practice, scorers often are required to obtain a specific level of agreement with validity papers (i.e., prescored samples of student work). If scorers fail to obtain the required level of agreement, they may be retrained or dropped from the scoring project. Once scorers have attained the qualification criteria, their work is typically monitored by scoring leaders who either rescore student work to check for scorer accuracy or circulate prescored validity papers. This is a time-consuming and labor-intensive task, requiring considerable time from not only scoring leaders, but also from scorers. It may be possible to use think-aloud methods as an alternative to current scorer-monitoring practices, reducing the amount of time individual scorers spend on qualifying and validity papers. It would be interesting to investigate the validity of such an alternative in future research.

Finally, our research has focused solely on essay scoring in a psychometric scoring system and, more specifically, proficiency within that system as defined by indices of interrater agreement. It is likely that the characteristics associated with scoring proficiency in such a context are quite different than are those associated with

proficiency in other contexts. In a psychometric scoring system, scorers are brought together for the purpose of learning how to apply an externally generated scoring rubric to the evaluation of student writing. There is little or no negotiation regarding the focus of this rubric. Scorers are expected to set aside their personal values and adopt those espoused by the rubric that they are trained to use. Scorers are rewarded subtly for their ability to think about writing like test developers, to come quickly to agreement with the scores assigned by test developers and other scorers, and to score essays quickly while maintaining reasonable levels of agreement.

These characteristics are probably very different than the traits that would describe scorers who would be outstanding in a hermeneutic scoring system (Moss, 1994). In such a system, scorers typically engage in social mediation as part of the evaluative process. Scorers integrate various perspectives and ancillary pieces of information about the student in an attempt to build the most comprehensive possible depiction of the student's performance. It is difficult to imagine how the characteristics that would make scorers successful in a psychometric scoring context would also be beneficial in a hermeneutic setting. Thus, we caution against overinterpreting the results of our research. Not only is it important to identify the characteristics that are associated with proficiency in the domain of essay scoring, but it is also important to identify the contexts in which those characteristics are likely to be beneficial to the individual. Additional research may be useful in determining how generalizable our results are to other scoring models.

APPENDIX A
Example Coded Protocol

Protocol	Scoring-Focus Code	Specificity Code	Rubric-Adoption Code	Processing-Action Code
Reads lines 1-5: "This first sentences isn't a complete sentence." "They use some strange words."	Mechanics	Specific	Self	Monitor
Reads lines 5-15: "Verb tense is incorrect here."	Style	Specific	Rubric	Monitor
Reads line 16 to end: "It's very difficulty to follow."	Mechanics	Specific	Self	Monitor
"The ideas ramble."	Storytelling	General	Rubric	Review
"There are lots of sentence fragments."	Organization	General	Self	Review
"There is some interesting vocabulary, like"	Mechanics	General	Self	Review
Recites section of text: Assigns score of 2 "It's descriptive."	Style	Specific	Rubric	Review
"But it isn't very clear."	Storytelling	General	Rubric	Rationale
"The writer should have organized the ideas better."	Storytelling Organization	General General	Rubric Rubric	Rationale Diagnose

Note: For this protocol, there are 10 coded statements—.30 of the statements are coded as belonging to the mechanics category for scoring focus, .60 of the statements are coded as belonging to the general category for specificity, .40 of the statements are coded as belonging to the self category for rubric adoption, and .20 of the statements are coded as belonging to the rationale category for processing actions.

Appendix B
Scoring Rubric

1. Narrative or storytelling—whether real or fictional—is one of the most familiar uses of language. At even this most basic level, the writer is likely to show evidence of understanding the nature of storytelling. The level of detail and specificity may be so minimal, however, that the reader has only the vaguest sense of action, or characters, or place. The story's structure may have essentially no discernible shape or direction. Vocabulary and sentence structures may be so simple that virtually no voice or style emerges. Control of surface features such as spelling,

capitalization, and usage may be so minimal as to make the language nearly incomprehensible, severely obstructing understanding.
2. At this level, it is clear that the writer intends to tell a story. That story may show very little development, but rudimentary detailing of action or character or place will appear. There is minimal evidence of a controlling structure for the story. Vocabulary and sentence structure may be extremely simple, and a personal voice may be identifiable. Lack of control of mechanical features may make it very difficult to follow the writer's ideas.
3. This level may be characterized as simple storytelling. Traditional narrative elements—such as character, setting, and action—are employed, but without much elaboration or sophistication. A clear although simple structure may be apparent; focus may shift or the writing may ramble. More varied vocabulary or sentence structure contributes to an emerging personal voice. The number or pattern of mechanical errors may make it difficult to follow the writer's ideas.
4. Writing samples at this level offer competent examples of storytelling. Narrative elements such as characterization, setting, action, or dialogue are employed with some skill; some detail and development will be apparent. The structure of sequence of the story supports the other elements adequately. Word choice and sentence structure are interesting and lend strength to the personal voice. Although there may be a number of mechanical errors, they seldom seriously interfere with understanding the writer's ideas.
5. Stories representing this level are engaging and interesting. The writer shows command of narrative elements, employing characterization, setting, action, and/or dialogue with skill; detail is specific and clear. The story's structure is strong and coherent. Word choice is precise and pleasing, and sentence structure is well-controlled, leading to a clear personal voice. Those mechanical errors that may be present do not often or seriously interfere with understanding the writer's ideas.
6. Stories representing this highest level not only are engaging and interesting, but also show some sophistication in thought, development, and/or word choice; they might be called memorable. The writer's command of narrative elements is evident in an engaging and sophisticated story. The story's structure is coherent, perhaps with an element of surprise. Word choice and sentence structure reveal a strong and readily identifiable individual voice. Although there may be some errors in mechanics, those errors do not noticeably interfere with understanding the writer's ideas.

Not ratable. These writing samples are unratable because they are blank, are completely illegible, are written in a language other than English, or entirely disregard the writing prompt.

SOURCE: American College Testing Program (1994).

REFERENCES

American College Testing Program. (1994). *Local scoring guide: 10th Grade Writing Assessment*. Iowa City, IA: Author.

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: Evaluating assessment essays. *College Composition and Communications, 37*, 315-327.

Breland, H. M. (1983). *The direct assessment of writing skill: A measurement review* (College Board Report No. 83-6). New York: College Entrance Examination Board.

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

Brennan, R. L. (1992). *Elements of generalizability theory* (2nd ed.). Iowa City, IA: ACT.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*, 65-81.

Cohen, J. (1988). *Statistical power analysis* (2nd ed.). 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.

Engelhard, G. J. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Fowles, M. E. (1978). *Basic skills assessment: Manual for scoring the writing sample. Analytic scoring/holistic scoring*. 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., & 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: Longman.

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.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

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.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). New York: Brroks/Cole.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of mathematics performance assessment. *Journal of Educational Measurement, 33*, 71-92.

Liebetrau, A. M. (1983). *Measures of association*. Newbury Park, CA: Sage.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345.

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.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

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.

White, E. M. (1985). *Teaching and assessing writing*. San Francisco, CA: Jossey Bass.

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

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*, 83-106.

Wolfe, E. W., & Feltovich, B. (1994, April). *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., Gitomer, D., & Carter, R. (1998, April). *The influence of changes in assessment design on the psychometric qualities of scores*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Edward W. Wolfe is currently an assistant professor in the Research and Evaluation Methodology Program in the College of Education at the University of Florida, where he specializes in measurement theory and statistics. Dr. Wolfe received his Ph.D. from the University of California, Berkeley, and prior to joining the faculty at the University of Florida, he spent 1 year as a postdoctoral fellow at Educational Testing Service in Princeton, New Jersey. Previously, he was employed as a program associate by ACT, Incorporated in Iowa City, Iowa. Dr. Wolfe's current research focuses on analysis of ratings and the rating process, particularly in the areas of performance-based and portfolio assessments.

Chi-wen Kao is currently a measurement statistician in the Kentucky Department of Education. Her specialization is educational measurement. She received her Ph.D. in the area of educational research from the University of Virginia.

Michael Ranney is both associate professor and head graduate advisor of the Graduate School of Education at the University of California, Berkeley, where he is also on the faculties of psychology, cognitive science, and math/science education (SESAME). A cognitive/experimental psychologist by training with a Ph.D. from the University of Pittsburgh, Ranney's research focuses on the relative coherence of human reasoning in diverse domains, especially those related to science. Much of his work also involves computational modeling, some of which he began during his postdoctoral work at Princeton University's Cognitive Science Laboratory. Ranney is currently studying the

nature of evidence and hypotheses and how they comprise explanations and contradictions in "Convince Me," a computational reasoner's workbench that his reasoning group has been developing over the past 5 years.