Journal of Writing Assessment Vol. 2, No. 1, pp. 37-56

Uncovering Rater's Cognitive Processing and Focus Using Think-Aloud Protocols

EDWARD W. WOLFE
Virginia Tech

This article summarizes the findings of a series of studies that attempt to document cognitive differences between raters who rate essays in psychometric, large-scale direct writing assessment settings. The findings from these studies reveal differences in both what information the rater considers as well as how that information is processed. Examining raters according to their ability to agree on identical scores for the same papers, this article demonstrates that raters who exhibit different levels of agreement in a psychometric scoring system approach the decision-making task differently and consider different aspects of the essay when making that decision. The research summarized here is an initial step in understanding the relationship between rater cognition and performance. It is possible that future research will enable us to better understand how these differences in rater cognition come about so that those who administer rating projects will be better equipped to plan, manage, and improve the processes of rater selection, training, and evaluation.

Performance and direct writing assessment has become more commonplace in large-scale assessment programs; and developers, researchers, and policymakers who work with these types of assessments have become increasingly aware of and attentive to the potential impact that rater effects may have on the validity and reliability of measures that are generated from these free-response,

Edward W. Wolfe is an associate professor of educational research and evaluation at Virginia Tech. Dr. Wolfe's research focuses on applications of Rasch models to instrument development and the analysis of ratings, influences of technology in testing on examinee mental states, and differential item functioning evoked by test translation.

Direct all correspondence to: Edward W. Wolfe, Educational Research and Evaluation, Virginia Tech, 313 East Eggleston Hall, Blacksburg, VA 24061, edwolfe@ct.edu

rated assessment instruments. This is particularly true when these measures are generated via a holistic scoring process. In an analytic scoring framework, a great deal of diagnostic information is communicated in the multiple measures generated for each of the various components of performance being evaluated. In a holistic scoring framework, on the other hand, the complexity of the written product or the complexity of creating a formulaic description of how the various components of a written product interact may preclude the generation of measures describing various aspects of performance. As a result, the meaning of the measures generated via holistic scoring frameworks may be more difficult to communicate to users of the assessment outcomes.

As a result, large-scale assessment programs in which holistic measures have implicit or explicit consequences for teachers or students lead administrators to seek holistic measures that exhibit high levels of reliability and validity for the sake of the legal defensibility and public acceptance of such measures. In contexts such as these, a common goal of the rater training and evaluation process is to develop a rating environment and community of raters in which raters think about and interpret student performance in similar ways under the assumption that raters, and hence the resulting ratings, are interchangeable. That is, a primary goal of those who direct rating sessions for large-scale performance and direct writing assessments is to ensure that raters think similarly enough about what constitutes a high- or low-quality student response that it does not matter which rater rates a particular response—the raters will assign interchangeable scores.

In large-scale, high-stakes assessment settings in which holistic scores are assigned, numerous practices have been adopted in an effort to minimize disagreements that arise due to the subjective nature of human judgments. For example, raters may be initially over trained (Kazdin, 1982), provided with testing and retraining as necessary (Medley, 1982), periodically recalibrated using benchmarks rated by expert raters (Kazdin, 1977), provided with feedback concerning the accuracy and levels of agreement their ratings exhibit (Curran, Beck, Lorrivean, & Monti, 1980), or monitored by having an expert rescore a sample of examinee responses rated by the rater in question (Longabaugh, 1980). However, regardless of the effort put forth to control for rater effects in large-scale performance and direct writing assessments, one thing remains clear—individual differences in both the content raters focus on and the processes raters use when rating student responses persist.

Unfortunately, little effort has been made to determine how or why raters differ in their rating practices or how these differences impact measures created from performance and direct writing assessments. The purpose of this article is to summarize the results of a series of studies that attempt to document cognitive differences between raters who rate essays in large-scale direct writing assessment settings. The analysis of these studies constitute some theoretical and research bases justifying the use of psychometric scoring systems to make important decisions about student writers.

A Model of Rater Cognition

Elsewhere, I have outlined an information-processing model of rater cognition within a psychometrically oriented scoring system (Wolfe, 1995, 1997; Wolfe, Kao, & Ranney, 1998). It is useful to distinguish the context of a psychometric scoring system from scoring systems that I call *hermeneutic scoring systems*. Several studies of rater cognition, particularly those focusing on direct writing assessment, have focused on hermeneutic scoring systems in which differences between raters are valued and are seen as opportunities for developing a richer and deeper understanding of the nature of the student's writing, the instructional implications of the student's writing, and the various interpretations that the writing affords the reader. Rater agreement in these local scoring systems comes from the shared context readers have about students and the specific, local decisions being made. Hermeneutic scoring systems are common in direct writing assessment contexts in which the implications (i.e., stakes) of student performance are local and a primary purpose of the assessment is make specific curricular decisions reflecting the standards and values of local populations and the institutions that serve them. Although I do not review relevant studies of hermeneutic writing assessment in this article, I refer interested readers to work by Haswell (2001), Smith (1993), O'Neill (2003),Vaughan (1991) and Huot (1993).

Differences between raters in *psychometric* scoring systems, on the other hand, are seen as possible sources of error that can detract from both the validity and reliability of the decisions being made on behalf of the assessment. Psychometric scoring systems are used to make decisions about large numbers of individuals across various local, cultural, and institutional contexts. Differences between raters are seen as indicators that the raters have not completely adopted the predefined scoring rubric or may not be suitable to read in a psychometric scoring context. Psychometric scoring systems are more common in large-scale assessment settings in which the implications of student performance are great and a primary purpose of the assessment is to provide information to policymakers and administrators concerning accountability, promotion, and selection decisions. As a result, raters in a psychometric scoring system are provided with a pre-defined scoring rubric, are trained to apply that rubric to examples of student writing, and are typically required to demonstrate proficiency with that rubric prior to being permitted to operationally score student responses.

Processing Actions

The information-processing model of rater cognition that I proposed for psychometrically oriented scoring contexts (Wolfe, 1995, 1997; Wolfe et al., 1998) differentiates between two cognitive frameworks—a *framework of scoring* and a *framework of writing* (as shown in Fig. 1). In that model, the rater reads the text written by the student and creates a mental image of the text. Of course, the created text images may differ from one rater to another due to environmental and experiential differences among raters (Pula & Huot, 1993). The text image is created and a scoring decision is made through the performance of a series

of *processing actions* that constitute the framework of scoring. That is, the framework of scoring is a mental script of a series of procedures that can be performed while creating a mental image of the text and evaluating the quality of that mental image. For example, raters *read* the text in order to begin formulating the text image. While reading, the rater may *comment* in a nonevaluative manner about his or her personal reactions to the content of the text. While formulating an evaluation of the text image, the rater may *monitor* specific characteristics of the text to determine how well the text exemplifies criteria that are set forth in the scoring rubric. After reading the text, the rater may *review* the features that seemed most noteworthy and then make a *decision* about the score to assign. Frequently, raters will provide a justification for that score by providing a *rationale* for the assigned score through a mapping of the features of the text onto the criteria laid out in the scoring rubric, *diagnosing* how the text could be improved, or *comparing* the essay to other texts the reader has read. These processing actions parallel the procedures outlined in a model of rater cognition presented by Freedman (Freedman & Calfee, 1983).

Differences between raters with respect to how the framework of writing is manifested during a rating episode may suggest important differences with respect to rater proficiency. For example, raters who employ read–monitor–read–monitor sequences while evaluating an essay may not adequately capture the essence of the writing in the text image that they create because they fail to identify important connections between ideas contained in the writing. Similarly, the tendency to make personal comments about the essay that are not of an evaluative nature may indicate that the rater is distracted from the rating process.

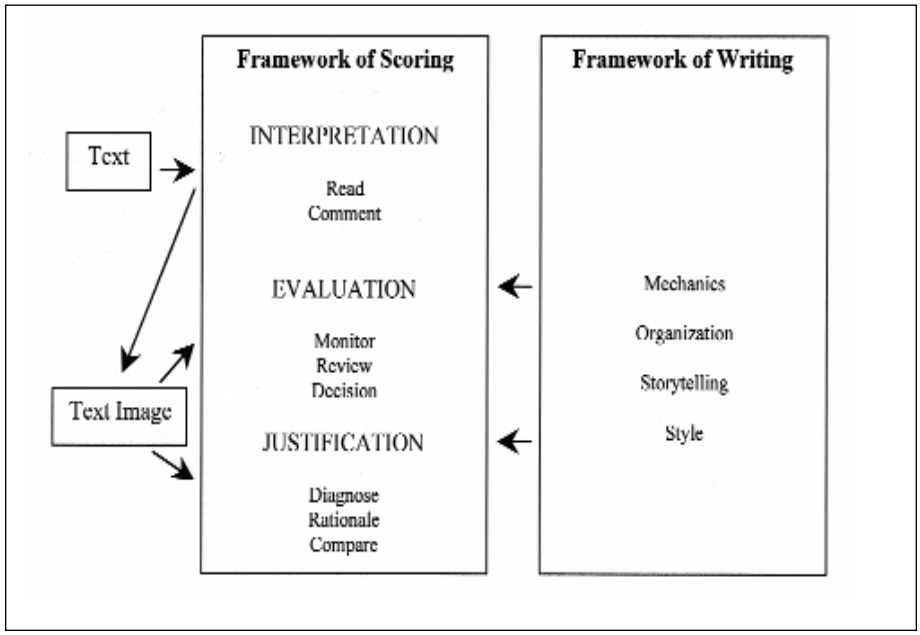

Fig. 1: An information-processing model of rater cognition.

Content Focus

The processes involved in creating an evaluation and rationale for the decision rely on the *framework of writing*—a mental image of the scoring rubric. The components of the framework of writing are developed based on the rater's individual experiences as well as efforts to train the rater to adopt the rubric designed by the test developers (Pula & Huot, 1993), and these components specify the characteristics of the student's writing that are important indicators of writing quality. As a result, a rater's framework of writing is likely to change from one rating project to another because different scoring rubrics are likely to be adopted for different projects (e.g., different expectations, different focus of assessment, different writing prompts or modes of writing). The specific writing characteristics that are incorporated into a rater's framework of writing for a particular scoring project are referred to here as the rater's *content focus*. For example, Fig. 1 displays key components for a framework of writing that are typical for a narrative writing task. Specifically, the rater's decisions are likely to be influenced by the quality of the *mechanics*, the *organization* of the student's ideas, the degree to which the student adopts *storytelling* devices to communicate the sequence of events, and the degree to which the student develops a unique *style* for presenting his or her ideas. These content focus categories parallel the notion of *interpretive frameworks* presented in a depiction of rater cognition presented by Frederiksen (1992).

Of course, raters may differ with respect to the way they define or understand the various components of the framework of writing. Differences in raters' explicit definitions of components of writing (e.g., mechanics, organization, etc.) were not considered in the research summarized here. However, two somewhat implicit aspects of their conceptualizations of the content focus categories were taken into account. The first aspect is the degree to which the rater makes general statements about the quality of a particular aspect of the writing and cites specific examples to illustrate the positive or negative qualities of the writing (referred to here as *specificity*). It may be that raters who focus on very specific aspects of the writing fail to build a holistic image of the text they have read. The second aspect is the degree to which raters utilize the vocabulary contained in the scoring rubric. The degree to which raters use words contained in the text of the scoring rubric, versus those self-generated by the rater (referred to here as degree of *rubric adoption*), may also have important implications for the degree to which the rater is able to come to consensus with other raters about the quality of a particular piece of writing—a trait that is highly valued in a psychometric scoring system.

In addition, important cognitive differences between raters may be revealed when raters place different emphases on the various content focus categories. For example, a rater who tends to consider the writer's style during an evaluation may come to a very different conclusion than a rater who places more emphasis on the use of storytelling devices. Not only may raters differ in the content focus that they adopt, but they may also differ with respect to the manner in which the various components of the framework of writing are considered. And, such differences in the manner in which the framework of writing manifests itself during a rating episode may reveal important differences between raters. For example, raters may

differ with respect to the number of content focus categories considered while evaluating a particular piece of writing (referred to here as *hits*). If a rater hits more categories, it may be evidence of a more thorough consideration of the various aspects of the essay while making a rating decision. Additionally, raters may differ with respect to the frequency with which they shift their focus between content focus categories while making a rating decision (referred to here as a *jump* from one category to another). A tendency to jump between categories more frequently may indicate a less principle-driven approach to evaluating an essay and may suggest that the rater is having trouble conceptualizing the writing as a whole rather than using the framework of writing as an organizational framework for approaching the rating process.

A Summary of the Model

Hence, the information-processing model depicts essay rating in a psychometric setting as a process of creating an image of the student's writing through the execution of a series of processing actions that constitute a framework of scoring. In addition, the rater relies on a framework of writing—a mental image of the scoring rubric that is created based on the rater's experiences as a writer, teacher, and rater as well as training that the rater undergoes as part of the current rating project (i.e., content focus). By executing a series of processing actions relating to evaluating the text image and providing a justification for a rating decision, the rater maps characteristics of the mental text image onto the mental image of the scoring rubric. Through this matching process, a best-fitting score is identified for the writing example in question. Various hypotheses can be generated concerning how individual differences in these processes and mental frameworks may manifest themselves as differences between the ratings assigned by raters in a psychometrically oriented rating system.

Data Collection

The remainder of this article summarizes a series of studies of rater cognition in large-scale direct writing assessments that employed a think-aloud methodology (Ericsson & Simon, 1993) in order to identify trends in the processing actions and content foci adopted by essay raters. Those studies focused on data obtained from 36 raters who rated 24 narrative writing samples written by 10th-graders for a large-scale direct writing assessment (Wolfe, 1995, 1997; Wolfe et al., 1998). These raters were selected from a pool of 60 raters. The pool of raters rated approximately 6,500 essays during a 1-week rating project, with each rater rating approximately 200 essays using a six-point rating scale created by the test developer (American College Testing, 1994).

Raters were initially trained to use the scoring rubric and then applied that scoring rubric to essays during the first two days of the rating session. At the end of the second day of rating, an intraclass correlation (Shrout & Fleiss, 1979) was computed for each of the raters in the rating pool. The intraclass correlation (r_{ic}) was computed to indicate the agreement between the ratings assigned to all essays that were

rated by an individual rater and the ratings assigned to these same essays by the randomly selected second raters. Three groups of participants (12 per group) were randomly selected 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 $r_{ic} = 0.74$. Intermediate raters showed relatively middle levels of agreement with other scorers with an average $r_{ic} = 0.80$. Proficient raters showed relatively high levels of agreement with other scorers with an average $r_{ic} = 0.87$.

For the think-aloud task, raters were asked to read 24 essays that were selected by a panel of writing assessment experts and test developers to represent a wide range of quality in student responses to the narrative prompts. In order to allow for privacy and to avoid creating a distraction during the rating project, the think-aloud task was conducted in a private room adjacent to the large room in which raters rated essays at various times during the days during which the rating project was conducted. Consistent with the guidelines described by Ericsson and Simon (1993), interviewers described the purpose and procedures of the study to the raters, presented the raters with the essays, and probed with questions like "remember to verbalize your thinking" only when raters spent more than a few seconds in silence. Responses were tape-recorded and were later transcribed to text so that content analyses could be performed.

Protocols were parsed into "thought units" (i.e., complete and independent thoughts). Then, by way of content analysis methods (Neuendorf, 2002), categories were developed through an iterative process in which initial coding categories were specified for processing actions and content codes as described in Wolfe (1995). Those categories were operationalized by defining the category and then identifying exemplars of the category through a review of the data. Category definitions were revised when examples were identified that could seemingly be coded into multiple categories or did not seem to adequately fit into any existing categories. When the iterative process failed to produce adjustments to the coding system, formal content analysis began. The appendix contains a summary of the content-coding categories that were developed, along with an example of a coded protocol.

Two individuals who had prior experience working with essay scorers as both essay scorers and as trainers of essay scorers performed formal coding and analyzed think-aloud protocols. Each parsed thought unit was coded according to six dimensions:

1. The essay feature being referenced (i.e., content focus).
2. The degree of specificity of the statement.
3. The degree of rubric adoption being demonstrated.
4. The number of content focus jumps.
5. The number of content focus hits.
6. The cognitive task being performed (i.e., processing action).

Each coder independently coded two thirds of the data so that both coders coded one third of the data. Cohen's kappa (κ) (Liebetrau, 1983) was computed for each

coding dimension, and intercoder agreement was deemed acceptable ($\kappa = 0.93$ for scoring focus, $\kappa = 0.87$ for degree of specificity, $\kappa = 0.91$ for degree of rubric adoption, and $\kappa = 0.85$ for processing actions).

The research questions are presented in Table 1.

Table 1. Hypotheses Investigated.

Characteristic	Coding Categories	Research Question
Processing actions	Comments	1. Do proficiency groups differ with respect to their rates of making personal comments about the essay?
Processing actions	Decision	2. Do proficiency groups differ with respect to their rates of making decisions about an essay prior to reading the entire essay?
Processing actions	Monitor, review, & rationale	3. Do proficiency groups differ with respect to their rates of using holistic versus atomistic approaches to rating essays?
Content focus	Jump	4. Do proficiency groups differ with respect to the rate with which they shift content focus categories during essay rating?
Content focus	Hit	5. Do proficiency groups differ with respect to the number of content focus categories mentioned when rating an essay?
Content focus	Content foci	6. Do proficiency groups differ with respect to the individual content focus categories that are cited across essays?
Content focus	Specificity	7. Do proficiency groups differ in the degree to which they cite specific characteristics versus make general references to the essay when describing essays?
Content focus	Rubric-adoption	8. Do proficiency groups differ in the degree to which they use rubric-based versus self-generated vocabulary when describing essays?

Differences Between Rater Proficiency Groups

Evidence of Processing Differences

Research Question 1 asks whether proficiency groups differ with respect to their rates of making personal comments about the essay. The rationale behind this question is that raters with different levels of proficiency with the rating task may exhibit different tendencies with respect to making connections with the writer while reading the text. Prior research relating to rater cognition does not provide a basis for predicting whether more or less proficient raters will make these connections, and two possible scenarios are readily apparent. First, it is possible that the more proficient raters, because they have automated the rating process, will be able to provide additional attention to nonevaluative details and would, therefore, be able to make more of these personal connections with the writer. Second, it is also possible that the less proficient raters, because they have not clearly formulated the rating task, would be more easily distracted from that task and, therefore, would be more likely to make these personal connections with the writer. However, data analysis provided no conclusive evidence of differences between the three groups with respect to rates of personal comments. As shown in Table 2, competent and intermediate raters tended to make only slightly more personal comments than did the proficient raters ($F[2,33] = 0.09$, $p = 0.91$).

Table 2: Descriptive Statistics for Personal Comments

	Proficient	Intermediate	Competent
Personal comments	23.42 (24.26)	27.17 (20.48)	25.92 (14.41)

Note. $n = 12$ for each group. Mean counts are shown with standard deviations in parentheses.

Research Question 2 inquired whether proficiency groups differ with respect to their rates of making decisions about an essay prior to reading the entire essay. Previous research concerning expertise in other decision-making domains suggests that experts are more likely to approach the decision-making process in a holistic, rather than atomistic, manner (Voss & Post, 1988). In the context of rater cognition, one might expect less proficient raters to approach the decision-making task by breaking it down into a series of smaller decisions, updating previous decisions once additional information was reviewed. Proficient raters, on the other hand, would be expected to review all of the available evidence prior to formulating an opinion. This would suggest that proficient raters would be less likely while competent raters would be more likely to make early decisions in their protocols (i.e., to voice a decision prior to reading the entire essay). In fact, the data indicate that proficient scorers made far fewer early decisions than did intermediate and compe-

tent raters. As shown in Table 3, intermediate and competent raters were much more likely to engage in early decision-making than more proficient raters, and the effect size is very large ($F[2,33] = 4.57$, $p = 0.02$, $d = 1.49$).

Table 3. Descriptive Statistics for Early Decisions

	Proficient	Intermediate	Competent
Early decisions	0.42 (0.90)	8.33 (9.69)	6.33(6.17)

Note. $n = 12$ for each group. Mean counts are shown with standard deviations in parentheses.

Research Question 3 also addressed the degree to which proficiency is related to adoption of a holistic versus atomistic approach to rating. And, consistent with the prediction made concerning Research Question 2, we would predict that the processing action use of proficient raters would be more consistent with a holistic approach and the processing action use of intermediate and competent raters would be more consistent with an atomistic approach. Specifically, one would expect proficient raters to use a read–then review–then decide sequence of processing actions and for intermediate and competent raters to use an iterative read–monitor–read–monitor–decide sequence. And, in fact, the data support this prediction. Specifically, proficient raters were more likely to use review processes while intermediate and competent raters were more likely to use monitor processing actions. Table 4 summarizes the proportion of evaluative comments each rater group made that fell into each processing action category. As predicted, proficient raters were much more likely to use review processing actions, whereas intermediate and competent raters were more likely to use monitor processing actions (monitor: $t[22] = 4.56$, $p = 0.0002$, $d = 0.88$; review: $t[22] = 3.49$, $p = 0.002$, $d = 0.68$).

Table 4. Descriptive Statistics for Processing Actions

Processing Action	Proficient	Intermediate	Competent
Monitor	0.06 (0.06)	0.31 (0.18)	0.24 (0.25)
Review	0.57 (0.18)	0.34 (0.14)	0.33 (0.27)
Diagnose	0.27 (0.13)	0.28 (0.12)	0.33 (0.25)
Rationale	0.10 (0.07)	0.07 (0.04)	0.10 (0.06)

Note. $n = 12$ for each group. Mean proportions are shown with standard deviations in parentheses.

Evidence of Content Focus Differences

Research Question 4 addressed the degree to which raters with different levels of proficiency differ with respect to their tendencies to shift attention during the evaluation process. Although previous research does not provide a basis for predicting the form of the relationship between rater proficiency and content focus category jump behaviors, it is easy to speculate what that relationship might look like. For example, we might expect the somewhat opportunistic nature of the read–monitor–read–monitor iterative reading style of intermediate and competent readers to lead to more frequent content focus category shifts by these raters. On the other hand, we might also expect these raters to exhibit a tendency to exhibit premature closure with respect to making a rating decision so that they shift their focus between content focus categories less often than do proficient raters. Regardless, the data shown in Table 5 provide no evidence of difference between proficiency groups with respect to jumping from one category to another ($F[2,33] = 0.31$, $p = 0.74$).

Table 5. Descriptive Statistics for Content Category Jumps

	Proficient	Intermediate	Competent
Jumps	0.69 (0.08)	0.69 (0.06)	0.67(0.80)

Note. $n = 12$ for each group. Mean proportions are shown with standard deviations in parentheses.

Similarly, Research Question 5 asked whether there are differences between proficiency groups with respect to the number of content focus categories mentioned in the think-aloud protocols. Again, although there is no basis in prior research to suggest that one proficiency group might be more comprehensive than another, it is easy to speculate that either group might be more prone to consider a greater number of categories while formulating a rating decision. It could be argued that proficient raters are more knowledgeable of the scoring rubric and that they attain their high levels of proficiency by focusing their energies on determining the degree to which the essay in question manifests each relevant characteristic outlined in that rubric. Alternatively, it could be argued that intermediate and competent raters, being more opportunistic in their approach to rating an essay, are more likely to hit all of the content focus categories in a particular evaluation because they discuss characteristics of the essay as they appear in the text. Proficient raters, on the other hand, may be more likely only to point out the key content focus categories that played into their rating decision because they approach the essay rating task from a more holistic perspective. But, again, the data, shown in Table 6, provide no evidence of group differences with respect to number of categories hit ($F[2,33] = 0.22$, $p = .81$).

Table 6. Descriptive Statistics for Content Category Hits

	Proficient	Intermediate	Competent
Hits	3.63(0.99)	3.48(0.73)	3.59(0.75)

Note. n = 12 for each group. Mean counts are shown with standard deviations in parentheses.

Research Question 6 asked whether there are differences between proficiency groups with respect to the relative frequency with which individual content focus categories are cited. Again, there is no basis in prior research to suggest which content categories each proficiency group might be more prone to cite while formulating a rating decision, but it is likely that the differences between raters with respect to their professional, teaching, and writing experiences would lead them to consider different aspects of the essay to be more or less important than would other raters with different experiences. Table 7 reveals that competent raters tended to place heavier emphasis on storytelling than did intermediate and proficient raters with a moderate effect size ($t[22] = 2.39$, $p = 0.03$, $d = 0.41$).

Table 7. Descriptive Statistics for Content Focus Categories

Scoring Focus	Proficient	Intermediate	Competent
Mechanics	0.12 (0.05)	0.08 (0.07)	0.13 (0.05)
Organization	0.23 (0.09)	0.32 (0.13)	0.20 (0.06)
Storytelling	0.44 (0.09)	0.41 (0.11)	0.50 (0.07)
Style	0.20 (0.06)	0.19 (0.08)	0.17 (0.05)

Note. n = 12 for each group. Mean proportions are shown with standard deviations in parentheses.

Research Question 7 focused on whether proficiency groups differ in the degree to which they cite specific characteristics of rather than make general references to the essay when making evaluative comments. Again, the tendency for less proficient raters to approach the task of rating an essay in an atomistic way would lead us to expect those raters to cite specific characteristics and more proficient raters to make more general comments. In fact, Table 8 supports this notion. This table reveals that competent raters made more specific comments than did intermediate and proficient raters. The effect size was moderate in magnitude ($t[22] = 2.39$, $p = 0.03$, $d = 0.41$).

Table 8. Descriptive Statistics for Degree of Specificity

Degree of Specificity	Proficient	Intermediate	Competent
General References	0.83 (0.10)	0.82 (0.07)	0.73 (0.11)
Specific Citations	0.17 (0.10)	0.18 (0.07)	0.27 (0.11)

Note. n = 12 for each group. Mean proportions are shown with standard deviations in parentheses.

The final research question asked whether proficiency groups differ in the degree to which they use rubric-based versus self-generated vocabulary when describing essays. Common sense suggests that proficient raters, as defined in a psychometric scoring system (i.e., one in which interrater agreement is valued), would be more likely to have internalized the scoring rubric, which would be exhibited by their tendency to use language generated from that rubric. Less proficient raters, on the other hand, would be more likely to use language that is not contained in the scoring rubric. Table 9 supports this notion. Specifically, the evaluative comments of proficient raters contained rubric-generated language about half of the time, whereas the language of intermediate and competent raters was self-generated about two thirds of the time. This difference is both statistically significant and meaningfully large (t[22] = 2.75, p = 0.01, d = 0.50).

Table 9. Descriptive Statistics for Degree of Rubric Adoption

Degree of Rubric Adoption	Proficient	Intermediate	Competent
Rubric-centered	0.47 (0.10)	0.34 (0.13)	0.34 (0.12)
Self-generated	0.53 (0.10)	0.66 (0.13)	0.66 (0.12)

Note. n = 12 for each group. Mean proportions are shown with standard deviations in parentheses.

Conclusions

An analyses of these studies reveals several interesting differences between raters to exhibit different levels of proficiency within a psychometric scoring system. Specifically, the results reveal differences in both what information the rater considers as well as how that information is processed. To summarize, we can say that raters who exhibit lower levels of proficiency within this

psychometric scoring system are more likely to focus on the ability of the writer to communicate a coherent story while the more proficient raters are more likely to consider the various characteristics of the essay equally. In addition, less proficient raters within this psychometric scoring system are also more likely to focus on very specific aspects of the essay and to rely on self-generated vocabulary when describing their thinking. Conversely, more proficient raters discuss the essay in more general terms using vocabulary that is contained in the scoring rubric adopted for the scoring project.

With respect to information processing differences between raters of different levels of proficiency, we see a similar trend. Specifically, less competent raters approach the rating task as a series of iterative decisions—a result that is consistent with previous research (Pula & Huot, 1993). First, the rater reads a short section of the essay and begins to formulate a decision about the quality of the essay. Next, the rater continues to read the essay and updates that decision frequently as additional information relevant to the scoring decision is encountered. On the other hand, a more proficient rater tends to read the entire essay withholding judgment until the entire essay has been read. This is evidenced by the fact that the less proficient raters in this study employed more early decisions and monitoring behaviors, whereas the more proficient raters employed review behaviors.

These differences are similar to differences between experts and novices in other fields (Glaser & Chi, 1988). In general, experts have been found to use top–down approaches to solving problems. That is, initially they spend a considerable amount of time thinking about the type of problem to be solved. After conceptualizing the type of problem that has been presented, the expert solves the problem quickly and accurately. This tendency parallels the observation in the studies reported here that experts tend to read the entire essay and reserve their evaluations until having completed the reading. It has been hypothesized that experts are able to manage the large volumes of information because they have created complex and interconnected networks for thinking about the domain in question through years of practice within that domain. As a result, experts seem to have attention made available for engaging in metacognitive thinking. For example, proficient raters (a) were more able to focus on a wider variety of content focus categories while making decisions, (b) utilized vocabulary contained in the scoring rubric, and (c) focused on general trends within the text rather than specific examples of weaknesses in the writing.

It is encouraging that these results jibe with studies of expertise in other domains, but much is yet to be understood with respect to the relationship between rater cognition and rater proficiency. For example, it is unclear how our knowledge of the interplay between prior knowledge and experiences should impact decisions about selecting the best pool of raters for a particular rating project, whether rater training procedures should be altered to take into account the cognitive characteristics of various raters, and how rater monitoring efforts might be improved as a result our understanding of rater cognition. In addition, it is important to understand how rater cognition varies across rating contexts. The context within which ratings were assigned in this study is very different from one in which ratings are assigned for a district writing assessment or by teachers within a school, and the results of this study would be difficult to apply to those quite different contexts.

The research summarized here is an initial step toward understanding the relationship between rater cognition and rater performance. This study demonstrates that raters who exhibit different levels of agreement in a psychometric scoring system approach the decision making task differently and consider different aspects of the essay when making that decision. It is possible that future research will enable us to better understand how these differences in rater cognition come about so that those who administer rating projects will be better equipped to plan, manage, and improve the processes of rater selection, training, and evaluation.

Appendix: Coding Categories and Coded Example Protocol

The following three sections outline the coding system that was developed and utilized in the studies that are summarized here. The first section describes the types of processing action codes that were assigned to parsed think-aloud statements. The second section describes the content focus codes. The third section presents an example of a coded protocol.

Processing Action Categories

Coding Category	Contexts	Definition
Comments	—	Nonevaluative comments about the contents of an essay, often relaying personal reactions of the rater to the message presented by the writer (e.g., "I like the way this person thinks!")
Decision	Early or late	A declaration of a score to be assigned to an essay. Early decisions occur before the rater has read the entire essay. Late decisions occur after the rater has read the entire essay (e.g., "This essay deserves a 4.")
Monitor	—	Evaluative comments concerning the characteristics of the essay that are presented by interrupting the reading process. The comments relate to content focus categories and are presented as evidence being considered while formulating a scoring decision (e.g., "I'm getting lost in the organization here—it's not very easy to follow."). Such comments are assumed to indicate that the rater is breaking down the rating process into smaller steps. As a result, a predominant use of monitoring processing actions is referred to here as an *atomistic* approach to rating.

Review	--	Evaluative comments concerning the characteristics of the essay that are presented after completion of the reading process but prior to assigning a score. The comments relate to content focus categories and are presented as evidence being considered while formulating a scoring decision (e.g., "I'm thinking that the organization isn't strong enough to support a 4."). Such comments are assumed to indicate that the rater has already formulated a complete text image. As a result, a predominant use of review and rationale processing actions is referred to here as a *holistic* approach to rating.
Rationale	--	Evaluative comments concerning the characteristics of the essay that are presented after assigning a score. The comments relate to content focus categories and are presented as a justification for a scoring decision (e.g., "I assigned a 3 because the organization was confusing in places."). Such comments are assumed to indicate that the rater has already formulated a complete text image. As a result, a predominant use of rationale and review processing actions is referred to here as a *holistic* approach to rating.

Content Focus Categories

Coding Category	**Contexts**	**Definition**
Content foci	Mechanics Organization Storytelling Style	Features or characteristics of an essay that are considered in an evaluation. In narrative writing common content foci include the mechanics (e.g., spelling, punctuation, grammar), organization (e.g., flow of ideas and connections between ideas via use of paragraphs and transitions), storytelling (e.g., the use of narrative devices to facilitate the communication of the sequence of events and main ideas), and style (e.g., the writer's personal voice as made evident by the use of vocabulary and tone).
Jump	--	A shift of attention from one content focus category to another.
Hit	--	The number of content focus categories mentioned during a particular evaluation.

Specificity	Specific General	The degree to which comments in which content focus categories mentioned contain references to specific features of the essay (e.g., a particular misspelled word, an awkwardly worded sentence, or an appropriate use of vocabulary to convey an emotion) or general trends in the essay (e.g., reference to the fact that some nonspecific words were misspelled, mentioning that the essay contained some awkward sentence structures, or a statement that the student used sophisticated vocabulary).
Rubric-adoption	Rubric Self	The degree to which comments in the content focus categories mentioned contain vocabulary that is contained in the scoring rubric versus vocabulary that is likely generated by the rater (self).

Example Protocol

What follows is an example protocol that has been coded according to the various codes mentioned in the previous two tables. The first column provides a quote of the rater's comment with vocabulary relevant to determining rubric-adopted versus self-generated content. The second column identifies the processing action being performed. The third column identifies the content focus of evaluative comments, the number of content category hits (*number), and the number of content category jumps (§number). The fourth column displays whether the comment references a general versus specific characteristic of the essay and whether the comment is rubric-adopted or self-generated in nature.

Hence, we can see that the rater begins by making a statement prior to reading the essay (monitor) about the organization (the first content focus hit) and that this statement is general in nature (i.e., does not reference a specific example of the characteristic). In addition, because the use of paragraphing was not referenced in the scoring rubric, the comment is self-generated.

In the next line of the table, the rater reads the first three lines of the essay and comments that "conscience" is misspelled. This constitutes a case of monitoring the content focus of mechanics. Hence, this constitutes the first jump (from organization to mechanics) and the second content focus category hit. In addition, this is a specific reference (i.e., identifies a particular example of poor mechanics), and it is rubric-generated because the rubric mentioned misspelling as a feature relating to mechanics.

In the third line of the table, the rater makes a nonevaluative comment, indicating a feeling of connecting with the writer. The table continues in this manner, indicating the processing action, content focus, hits, jumps, specificity, and rubric adoption of the comments.

Quote	Processing Action	Content Focus	Other
Before reading anything, it is clear that the student didn't use paragraphs.	Monitor	Organization*1	General Self
[READS FIRST 3 SENTENCES.] "Conscience" is misspelled in that sentence.	Monitor	Mechanics*2 §1	Specific Rubric
[READS NEXT 3 SENTENCES] This student reminds me of my bratty nephew. He always whines when he doesn't get his way.	Comment		
[READS NEXT 2 SENTENCES] OK. Even though the student doesn't use paragraphs, there are some nice transitions between ideas here.	Monitor	Organization§2	General Self
The student is doing a good job of providing the essential details for understanding why the event is so important.	Monitor	Storytelling*3§3	General Rubic
This essay will probably get a 4.	Decision (Early)		
[READS REMAINDER OF ESSAY] Well, the story gets told. I have a clear picture of what happened.	Review	Storytelling	General Self
And, the ideas seem to flow pretty well.	Review	Organization§4	General Rubric
But, the writer also adds a bit of a personal touch to the story by using a unique voice for telling the story.	Review	Style*4 §5	General Self/Rubic
Well, the spelling, punctuation, and grammar aren't very good.	Review	Mechanics§6	General Rubric
I'd give this a 4.	Decision (Late)		
The mechanics are too weak to warrant a 5.	Rationale	Mechanics	General

References

American College Testing Program. (1994). *Local scoring guide: 10th grade writing assessment.* Iowa City, IA: Author.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Logan: Utah State University Press.

Curran, J. P., Beck, J. G., Corriveau, D. P., & Monti, P. M. (1980). Recalibration of raters to criterion: A methodological note for social skills research. *Behavioral Assessment, 2*, 261-268.

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

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.

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: Erlbaum.

Haswell, R. (Ed.). (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Westport, CT: Ablex.

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.

Kazdin, A. E. (1977). Artifact, bias, and complexity: The ABC's of reliability. *Journal of Applied Behavior Analysis, 10*, 141-150.

Kazdin, A. E. (1982). Observer effects: Reactivity of direct observation. In D. P. Hartmann (Ed.), *Using observers to study behavior* (pp. 5-19). San Francisco, CA: Jossey-Bass.

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

Longabaugh, R. (1980). The systematic observation of behavior in naturalistic settings. In H. C. T. J. W. Berry (Ed.), *Handbook of cross-cultural psychology: Vol. 2. Methodology* (pp.). Boston: Allyn & Bacon.

Medley, D. M. (1982). Systematic observation. In H. E. Mitzel (Ed.), *Encyclopedia of educational research* (5th ed., pp. 1841-1851). New York: Macmillan.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

O'Neill, P. (2003). Moving beyond holistic scoring through validity inquiry. *Journal of Writing Assessment, 1,* 47-65.

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Smith, W. L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M.M. Williamson & B.A. Huot (Eds.) *Validating holistic scoring for writing assessment* (pp. 142-205). Cresskill, NJ: Hampton Press.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 11-125). 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: Erlbaum.

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., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication, 15*, 465-492.