Expertise in Essay Scoring

Edward W. Wolfe
Performance Assessment Center
American College Testing
USA
wolfe@act.org

Michael Ranney
EMST
Graduate School of Education
University of California, Berkeley
USA
ranney@cogsci.berkeley.edu

Achieving levels of reliability that allow large-scale essay assessments to be used to guide educational policy is a major hurdle for test developers. Previous studies have shown that one influential source of measurement error associated with essay scores is rater idiosyncracies [Engelhard 1994]. Although the literature dealing with scoring cognition is not conclusive about why some scorers are more consistent than others, it offers some insight into variables that may account for individual differences in scoring competence [Pula & Huot 1993; Wolfe & Feltovich 1994].

The purpose of this study is to empirically identify characteristics that differentiate essay scorers of differing levels of proficiency. By identifying the characteristics that differentiate better from poorer scorers, developers of large-scale essay examinations may be able to make scorer training and monitoring efforts less costly and more effective.

Theoretical Framework

Based on the literature on scoring cognition [Freedman & Calfee 1983; Huot 1993; Vaughan 1991; Wolfe & Feltovich 1994], we developed a model of scoring cognition that contains two components: processing actions and content focus. Processing actions are the mental operations that are used to compare an essay's content to a set of scoring criteria. Here is an example of how processing actions are applied during the scoring process: As they read an essay, scorers occasionally interrupt their reading to monitor how well the essay satisfies the scoring criteria. After reading the paper, scorers may review the essay to note its strengths and weaknesses. They may also diagnose problems with the essay and suggest ways to improve it. Finally, the scorer assigns a score and provides a rationale for that score.

We define content focus as the features of the essay upon which scoring decisions are based. The content focus for this study includes the ability to tell a story, essay organization, individual writing style, and control of mechanics as well as other less relevant features of the essay (e.g., textual appearance, how well the writing addresses the assignment, and non-specific comments).

Based on findings in similar domains of expertise, one might expect the knowledge structures of experts to be organized in a more sophisticated way, allowing them to perceive information in the form of large meaningful patterns and to access this information more quickly and with deeper understandings than novices [Voss & Post 1988]. Prior research on scorer cognition led us to formulate four hypotheses about scoring expertise. Hypothesis 1 predicted that the more proficient scorers as a group would be more consistent in their use of processing actions. Hypothesis 2 predicted that the more proficient scorers as a group would be more consistent
in their use of *content focus* categories. Hypothesis 3 predicted that scorers of different proficiency levels would emphasize different *processing actions*. Hypothesis 4 predicted that scorers of different proficiency levels would emphasize different *content focus* categories.

**Method**

**Subjects**

Subjects for this study were 36 scorers who took part in a large essay scoring project. Based on their demonstrated levels of proficiency with the scoring rubric, subjects were selected to equally (i.e., by 12’s) represent three proficiency groups: *novices*, *intermediates*, and *experts*. Subjects performed a think aloud task as they scored 24 essays. Interviews were audiotaped and transcribed for analysis. Each statement made by a scorer was coded according to its *content focus* (i.e., appearance, assignment, mechanics, non-specific, organization, storytelling, or style) and its *processing action* (i.e., diagnose, monitor, review, or rationale). The proportions of statements that fell into each coding category across essays served as the data for making group comparisons. Cohen’s $\kappa$ was .85 and .93 for the *content focus* and *processing action* codes, respectively.

**Analyses**

Multiple $t$ tests were employed to investigate our four hypotheses. Our goal was to identify monotonic relationships between the cognitive habits of scorers and scoring proficiency. For Hypotheses 1 and 2, two a *priori* orthogonal contrasts were applied to the variances of the proportions for each *processing action* and *content focus* category. The two contrasts compared *experts* to the combined group of *intermediates* and *novices* ($\Psi_1$: $\sigma^2_{\text{expert}} - 1/2(\sigma^2_{\text{intermediate}} + \sigma^2_{\text{novice}})$) and *intermediates* to *novices* ($\Psi_2$: $\sigma^2_{\text{intermediate}} - \sigma^2_{\text{novice}}$). For Hypotheses 3 and 4, similar analyses were performed on the means of the proportions for each *processing action* and *content focus* category ($\Psi_1$: $\mu^2_{\text{expert}} - 1/2(\mu^2_{\text{intermediate}} + \mu^2_{\text{novice}})$ and $\Psi_2$: $\mu^2_{\text{intermediate}} - \mu^2_{\text{novice}}$). Post hoc analyses were performed on additional variables identified by protocol coders.

**Results**

Table 1 shows the variance of the *processing action* proportions for the think aloud data by proficiency group (Hypothesis 1). These data show that *experts* were more consistent in their use of *monitor* actions than were *intermediates* and *novices*; $t(33) = -4.21, p < .001$. Although the difference between *intermediates* and *novices* for monitoring was not statistically significant ($t(33) = 1.00, p = .16$), the data suggest that there may be an increase in the consistency with which *monitor* actions are used as scoring proficiency increases. The *intermediates* were significantly more consistent than *novices* for the *review* and *rationale* *processing actions*; $t(33) = -2.99, p = .004$ and $t(33) = -2.53, p = .008$, respectively. Although the contrasts did not afford a comparison of the differences between *experts* and *intermediates*, inspection of the data seems to indicate that the variances of *experts* and *intermediates* are similar for both of these variables [Table 1].

<table>
<thead>
<tr>
<th>Processing Action</th>
<th>Expert $\sigma^2$</th>
<th>Intermediate $\sigma^2$</th>
<th>Novice $\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitor</td>
<td>0.0034</td>
<td>0.0318</td>
<td>0.0649</td>
</tr>
<tr>
<td>Review</td>
<td>0.0326</td>
<td>0.0204</td>
<td>0.0717</td>
</tr>
<tr>
<td>Rationale</td>
<td>0.0182</td>
<td>0.0138</td>
<td>0.0636</td>
</tr>
<tr>
<td>Diagnose</td>
<td>0.0055</td>
<td>0.0013</td>
<td>0.0036</td>
</tr>
</tbody>
</table>

Table 1: Group Variances on Processing Actions
Similar analyses compared the variances of the proportions for the content focus codes (recall Hypothesis 2). Table 2 shows the group variance of the proportions for the content focus data. Only one of the contrasts in these data was statistically significant—experts were less variable in their use of storytelling than were intermediates and novices; \( t(33) = -2.18, p = .02 \). However, this difference cannot be easily interpreted from a learning perspective because the relationship between scoring proficiency and storytelling use is non-monotonic. Furthermore, most of these trends are opposite those predicted by Hypothesis 2. That is, scorers who were more proficient were typically less consistent in their use of content focus categories than were less proficient scorers [Table 2].

<table>
<thead>
<tr>
<th>Content Focus</th>
<th>Expert ( \sigma^2 )</th>
<th>Intermediate ( \sigma^2 )</th>
<th>Novice ( \sigma^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appearance</td>
<td>0.0010</td>
<td>0.0038</td>
<td>0.0012</td>
</tr>
<tr>
<td>Assignment</td>
<td>0.0006</td>
<td>0.0004</td>
<td>0.0004</td>
</tr>
<tr>
<td>Mechanics</td>
<td>0.0016</td>
<td>0.0025</td>
<td>0.0019</td>
</tr>
<tr>
<td>Non-Specific</td>
<td>0.0064</td>
<td>0.0066</td>
<td>0.0024</td>
</tr>
<tr>
<td>Organization</td>
<td>0.0068</td>
<td>0.0094</td>
<td>0.0020</td>
</tr>
<tr>
<td>Storytelling</td>
<td>0.0024</td>
<td>0.0131</td>
<td>0.0043</td>
</tr>
<tr>
<td>Style</td>
<td>0.0026</td>
<td>0.0056</td>
<td>0.0018</td>
</tr>
</tbody>
</table>

Table 2: Group Variances on Content Focus

The analyses for Hypothesis 3 were based on the data in Table 3 which shows the descriptive statistics of the proportions for the processing action codes for each proficiency group. Two of the group comparisons were statistically significant. Experts were less likely to use monitor actions than were intermediates and novices; \( t(21) = -4.45, p = .001 \). On the other hand, experts were more likely to use review actions than were intermediate and novice scorers; \( t(33) = 3.23, p = .005 \). Statistical tests did not indicate any differences between intermediates and novices for any of these variables [Table 3].

<table>
<thead>
<tr>
<th>Processing Action</th>
<th>Expert</th>
<th>Intermediate</th>
<th>Novice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( M )</td>
<td>( SD )</td>
<td>( M )</td>
</tr>
<tr>
<td>Monitor</td>
<td>.06</td>
<td>.06</td>
<td>.31</td>
</tr>
<tr>
<td>Review</td>
<td>.57</td>
<td>.18</td>
<td>.34</td>
</tr>
<tr>
<td>Rationale</td>
<td>.27</td>
<td>.13</td>
<td>.28</td>
</tr>
<tr>
<td>Diagnose</td>
<td>.10</td>
<td>.07</td>
<td>.07</td>
</tr>
</tbody>
</table>

Table 3: Group Proportions on Think Aloud Processing Actions

Table 4 presents the descriptive statistics for the group proportions for the content focus codes (recall Hypothesis 4). Two of the group comparisons were statistically significant. Intermediates were more likely to make organization content statements than were novices; \( t(16) = 2.60, p = .01 \). However, as was true for the variance comparison for organization, a visual inspection of the three group means renders this finding uninterpretable in a learning context. The second statistically significant difference revealed that intermediate scorers were less likely to make storytelling comments than were novices; \( t(33) = -2.74, p = .01 \). Experts were similar to intermediates in their use of this content focus category as shown by the data [Table 4].

547
Based on observations of the protocol coders, post hoc analyses were performed on group differences regarding the extent to which scorers tended to withhold judgment during the decision making process. An examination of the think aloud protocols suggested that experts were more likely to approach the scoring task as a two-phased process. First, the expert scorer would read the essay, typically from beginning to end, without interrupting the reading to comment on the essay’s content (i.e., without monitor comments). Second, the expert scorer would review the essay’s contents and announce a score for the essay. Some experts made very few comments about the essay’s content, seeming to arrive at an implicit decision rather than feeling the need to explicitly review the essay’s content or provide a rationale for the assigned score.

Novices and intermediates, on the other hand, seemed to be more likely to monitor the essay’s content, a process that interrupts the reading process. Some of the less proficient scorers even performed monitoring actions prior to reading any portion of the essay. These scorers took a cursory glance at the essay or read the first sentence of the essay and described what they expected the remainder of the essay to contain. These less proficient scorers were also more likely to interrupt the reading process to predict how they would eventually score the essay or to update a prior prediction. One of the novice scorers even assigned scores to most of the essays without reading the entire essay. Other scorers mentioned that they typically had a score in mind well before completing the essay, but none of them failed to read the entire essay during their think-aloud interviews.

To see whether these observations could be supported statistically, the number of early decisions for each scorer (i.e., scores that were announced during the reading of the essay) were summed across all essays. Table 5 shows the descriptive statistics for each proficiency group. Because these were post hoc analyses, Hartley’s $F_{\text{max}}$ test was used to test the homogeneity of variances assumption. This test revealed that the group variances were not equal, $F_{\text{max}} (3,11) = 115.90, p < .05$, so the groups were compared with a studentized range statistic ($q$) to test all pairwise comparisons, controlling type I errors at $\alpha = .05$. The differences between experts versus intermediates as well as the difference between experts versus novices were statistically significant ($q (11) = -3.99, p = .04$ and $q (11) = -5.92, p = .03$, respectively) and the difference between intermediates and novices was not; $q (19) = 0.85, p > .10$ [Table 5].
<table>
<thead>
<tr>
<th>Group</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experts</td>
<td>0.42</td>
<td>0.90</td>
</tr>
<tr>
<td>Intermediates</td>
<td>8.33</td>
<td>9.69</td>
</tr>
<tr>
<td>Novices</td>
<td>6.33</td>
<td>6.17</td>
</tr>
</tbody>
</table>

Table 5: Group Frequencies for Early Decisions

Discussion

These results are consistent with findings of similar studies of scorer cognition [Vaughan 1991; Huot 1993; Wolfe & Feltovich 1994]. However, our study suggests three trends in the development of scoring expertise that have not been offered by other researchers.

**Conclusion 1:** Experts are more consistent than non-experts in the way they approach the task of scoring essays. The data supporting Hypothesis 1 indicate that expert essay scorers are more likely to use similar strategies for scoring than are intermediates and novices. One explanation of the dissimilarities in the three groups' approaches to scoring may be that these processing differences are due to differences in the knowledge structures that underlie expert-like performance, leading experts to use that knowledge more effectively. Similar conclusions have been drawn from the results of numerous studies of expert performance in a variety of problem-solving domains.

**Conclusion 2:** Expert scorers are more likely to use more fluent methods of scoring essays than are non-experts. That is, the data supporting Hypothesis 3 suggest that expert scorers seem to utilize a more holistic strategy for scoring that uses a less iterative decision making pattern than that of non-experts. Experts seem to use strategies in which the scorer interprets the student writing through reading and reacting to the text, thus creating an image of the text. They then map the features of this text image onto the their mental representations of the scoring criteria. Through this process, judgments are made about how well the writer has demonstrated the various aspects of the scoring criteria, and a decision is formulated about the score to assign to the essay.

On the other hand, intermediate and novice scorers seem to use less fluent strategies for scoring. That is, non-experts seem to go through an alternating cycle of reading and monitoring portions of the essay. During each iteration, the scoreable features of that section of the essay are mapped onto the scorer's mental representation of the scoring criteria, and a preliminary score may be assigned to the essay. After completing this process for the entire essay, non-experts may review the essay prior to assigning a final score. However, they are less likely to do so than are experts.

This interpretation does not necessarily suggest that the approach taken by non-experts is inferior to that taken by experts. It is quite plausible that similar processing is occurring during the reading phase of an expert's scoring. However, more emphasis seems to be placed on the reading and comprehension process by experts than by intermediates and novices [Huot 1993]. It may be that experts have simply automated these procedures.

**Conclusion 3:** There is little evidence to suggest that the content focus adopted by essay scorers is related to scoring proficiency. Scorers in this study demonstrated similar emphases in their content foci as has been observed in other studies of scoring [Huot 1993; Vaughan 1991]. That is, primary attention has been given to storytelling, organization, and style. Unfortunately, the analyses associated with Hypotheses 2 and 4 provided little evidence that content focus has any relationship with scoring proficiency.

The most appealing explanation for the observed differences in processing action use by the proficiency groups is that these differences are caused by structural differences in the knowledge upon which those actions operate.
Given the myriad of studies of expertise in other domains of human performance that have indicated that the
primary difference between experts and non-experts lies in the manner in which domain knowledge is structured,
a possible explanation for the failure to detect differences in the content focus of the proficiency groups in this
study may be that the coding system used in this study was not sensitive enough to the kinds of differences in
knowledge structures that make expert-like processing possible. With hindsight, it seems unlikely that this
coding system could have identified such differences because it emphasizes the focus of the comments that a
scorer makes rather than the structure of the knowledge (i.e., the relationships between concepts) that underlies
those comments. Future studies using other means of analysis should aim to determine whether there are
differences in these knowledge structures.

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

References


and cognitive theory. In P. Mosenthal, L. Tamor, & S.A. Walmsley (Eds.), Research on writing: Principles and methods
(pp. 75-98). New York, NY: Longman.

Williamson & B.A. Huot (Eds.), Validating holistic scoring for writing assessment (pp. 206-236). Cresskill, NJ: Hampton
Press.

Williamson & B.A. Huot (Eds.), Validating holistic scoring for writing assessment: Theoretical and empirical foundations


& M.J. Farr (Eds.), The nature of expertise (pp. 261-285). Hillsdale, NJ: Lawrence Erlbaum.


Acknowledgements

This project was funded by American College Testing. The authors also thank Carol Liguori for assisting with the data
coding, Chi-wen Kao for assistance in data analysis, and Brian Feltovich and Florian Kaiser for reviewing the final draft
of this manuscript.