DOCUMENT RESUME

ED 396 005 TM 025 346

AUTHOR Wolfe, Edward W.; Kao, Chi-Wen
TITLE The Relationship between Scoring Procedures and Focus and the Reliability of Direct Writing Assessment Scores.
PUB DATE 8 Apr 96
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Content Analysis; Educational Assessment; *Essays; Evaluation Methods; *Evaluators; *Scoring; *Test Reliability; Test Use; *Writing Tests
IDENTIFIERS *Direct Assessment; Monitoring

ABSTRACT

This paper reports the results of an analysis of the relationship between scorer behaviors and score variability. Thirty-six essay scorers were interviewed and asked to perform a think-aloud task as they scored 24 essays. Each comment made by a scorer was coded according to its content focus (i.e. appearance, assignment, mechanics, communication, organization, story, or style) and its processing action (i.e. diagnose, monitor, review, or rationale). The number of comments made by each scorer that fell into each of these categories was regressed onto the variability of the scores assigned to each of the 24 essays. The results show that direct writing assessment scores are less reliable on essays for which scorers focus on abstract and difficult-to-define features. More specifically, these scorers showed less agreement for essays when their evaluative comments focused on the way the story is communicated by the writer and the way that the writer's sentence structure, vocabulary, and voice convey an individual's writing style. Scorers were also more likely to use less efficient scoring strategies for essays such as these. That is, they were more likely to break the scoring task down into subtasks by using a monitoring strategy. They were also more likely to diagnose ways that the writing could be improved. (Author)

Running head: Cognition & Reliability

The relationship between scoring procedures and focus and the reliability of direct writing assessment scores

Edward W. Wolfe

American College Testing, Iowa City, Iowa

Chi-Wen Kao

University of Virginia

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

Abstract

This paper reports the results of an analysis of the relationship between scorer behaviors and score variability. Thirty-six essay scorers were interviewed and asked to perform a think aloud task as they scored 24 essays. Each comment made by a scorer was coded according to its content focus (i.e., *appearance, assignment, mechanics, communication, organization, story,* or *style*) and its processing action (i.e., *diagnose, monitor, review,* or *rationale*). The number of comments made by each scorer that fell into each of these categories was regressed onto the variability of the scores assigned to each of the 24 essays. The results show that direct writing assessment scores are less reliable on essays for which scorers focus on abstract and difficult-to-define features. More specifically, these scorers showed less agreement for essays when their evaluative comments focused on the way the *story* is communicated by the writer and the way that the writer's sentence structure, vocabulary, and voice convey an individual's writing *style*. Scorers were also more likely to use less efficient scoring strategies for essays such as these. That is, they were more likely to break the scoring task down into subtasks by using a *monitoring* strategy. They were also more likely to *diagnose* ways that the writing could be improved.

The relationship between scoring procedures and focus and the reliability of direct writing assessment scores

Achieving levels of reliability that allow for large-scale comparisons and selection decision in a cost-effective manner has been a difficult task for those who develop direct writing assessments. Generalizability theory has provided test developers with methods of identifying variables that may influence the reliability of these assessments. However, suggestions for increasing test reliability based on the results of generalizability studies have been limited to either increasing the number of prompts to which a student responds or increasing the number of raters who score a student's responses (Linn, 1992)--both of these approaches result in increases in testing costs.

Other researchers have taken a different approach to identifying potential sources of construct irrelevant variance by studying how variables related to essay quality (e.g., handwriting quality, mechanical errors, opportunity for revision) or raters (e.g., expertise or gender) influence writing assessment scores. These researchers have provided valuable insights into qualitative variables that may influence test reliability. Unfortunately, researchers taking this approach have limited their focus to the mean difference between groups of examinees rather than investigating how these variables influence score variability.

The purpose of this study is to investigate the relationship between some of these qualitative variables and direct writing assessment score variance. More specifically, this study attempts to identify how the processes and focus adopted by an essay scorer are related to interrater reliability. By understanding how scoring methods influence score variability, testing organizations may be able to increase reliability without increasing testing costs.

Theoretical Framework

Vaughan (1991) performed some of the first research on the thinking processes used by essay scorers. This work identified a variety of scoring procedures (i.e., the methods used to read and review the essay) and trends in content focus (i.e., how essay characteristics such as organization or mechanics are considered during the decision making process) used by essay scorers. Further work in this area was done by Huot (1993), who examined differences in the content focus and scoring procedures used by novice and experienced scorers. Huot found that experienced scorers were more likely to read the essay with less disruptions and then comment on the essay after reading. Novice scorers, on the other hand, were more likely to read and score at the same time. Wolfe and Feltovich (1994) found similar differences between the scoring procedures used by expert and novice scorers. Their study also revealed that experts tended to focus their evaluative comments on more complex and abstract features of the essay (e.g., the writer's voice or ability to communicate ideas) while novices focused on more concrete features (e.g., mechanics or textual appearance).

In recent work in the area of performance assessment, Frederiksen (1992, April) suggested that teacher evaluators use interpretive frameworks to understand and evaluate teacher performance. Frederiksen's notion of an interpretive framework suggests that the evaluator monitors a performance for some set of criteria (defined by the evaluator's interpretive framework). When a noteworthy instance of a performance criteria occurs, the evaluator makes a mental note of the aspect of the criteria being demonstrated and the degree of competence shown at that instance. Thus, the interpretive framework serves as a means for understanding and recognizing the parameters of the performance being assessed. After all noteworthy moments have been observed, the evaluator considers all of the observations, weights them, and decides

on a score. The final step in the process is to create a rationale. Thus, the interpretive framework is also used to organize and communicate ideas about the performance to others.

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

In this model, information is taken from the printed text, and an image of the student response is constructed. The scorer *interprets* student writing based on his or her own world knowledge, beliefs and values, and knowledge of the writing process. Aspects of the reading environment may also influence the form that this text image takes. This means that the text image is not an exact replication of the original text and that one scorer's text image may be very different than text images constructed by other scorers. Based on the created image of the text, the scorer compares various aspects of the writing to representations of the scoring criteria. Through this process judgments are made about the text, and a decision is formulated about how well the writer has demonstrated competence in writing. Finally, the evaluative decision is articulated through written or oral comments about the text.

Wolfe and Feltovich (1994, April) and Wolfe (1995) extended this work by proposing a model of scorer cognition that allows for both variations in the features of an essay upon which scoring decisions are based as well as variations in the procedures that are used to make these

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

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

different scoring decisions for the same essay. This model of scoring cognition was adopted for the study reported in this paper, so the following sections provide a detailed description of how frameworks of writing, frameworks of scoring, and processing actions are manifested in the behaviors of essay scorers.

Insert Figure 2

Based on this model, two types of scorer behaviors were identified that might influence score variability: *processing actions* and *content focus*. The processing actions investigated in this study were *monitor, review, rationale*, and *diagnose*. Scorers *monitor* when they interrupt their reading of the essay to identify its strengths and weaknesses. *Reviewing*, on the other hand, occurs when a scorer has finished reading the essay and reexamines its content prior assigning a score. After a score has been assigned, a scorer may provide a *rationale* for that score by citing characteristics of the essay as a justification. Throughout the process, scorers may also *diagnose* ways that the essay could be improved. The *content focus* examined in this study were the writer's ability to *communicate* ideas, the writer's ability to tell a *story*, the essay's *organization*, individual writing *style*, control of *mechanics*, textual *appearance*, and how well the writing addresses the *assignment*.

Based on this model of scorer cognition, our study investigated two questions: *1) Is score variance related to the use of processing actions employed by scorers? and 2) Is score variance related to the content that scorers focus on while making scoring decisions?*

Method

Subjects for this study were 36 essay scorers who took part in a large-scale writing assessment scoring project. Subjects were interviewed individually and was asked to perform a think aloud task (i.e., to verbalize their thinking) as they scored 24 essays. The interviews were

recorded and then transcribed to written form for analysis. Each comment made by a scorer was coded according to its content focus (i.e., *appearance, assignment, mechanics, communication, organization, story,* or *style*) and its processing action (i.e., *diagnose, monitor, review,* or *rationale*). The number of comments across scorers that fell into each coding category for each of the 24 essays served as the independent variables for this study. The variance of the scores assigned to each essay served as the dependent variable.

This resulted in a data matrix containing the total number of comments made for each content focus category and for each processing action category for each essay as well as the variance of scores for that essay. Regression analysis was used to identify the content focus categories that were most strongly associated with score variance. Counts for *appearance, assignment, mechanics, communication, organization, story,* and *style* were regressed onto score variances using a forward entry method with tolerance set at .01. Separate analyses were performed on the processing action data.

Results

Table 1 shows that the best fitting model for content focus included *story* and *style*. Table 2 shows that the best fitting model for processing actions included *monitor* and *diagnose*.

Table 1 here

Table 2 here

These results show that both the content focus and processing actions adopted by scorers are related to the variability of the scores they produce. In terms of content focus, a high number of comments about *story* and *style* were correlated with higher score variances; F (2,21) = 6.78, p = .005, R^2 = .39. In terms of processing actions, both *monitor* and *diagnose* comments were positively associated with score variance; F (2,21) = 3.34, p = .05, R^2 = .25.

Discussion

This study shows that direct writing assessment scores are less reliable on essays for which scorers focus on rather abstract and difficult-to-define features. More specifically, these scorers showed less agreement for essays when their evaluative comments focused on the way the *story* is communicated by the writer and the way that the writer's sentence structure, vocabulary, and voice convey an individual's writing *style*. Scorers were also more likely to use less efficient scoring strategies for essays such as these. That is, they were more likely to break the scoring task down into subtasks by using a *monitoring* strategy. They were also more likely to *diagnose* ways that the writing could be improved. For essays with lower score variance, scorers were more likely to read the entire essay from beginning to end before making evaluative comments.

Future research related to these findings could take two directions. One approach would be to identify ways to improve scorers' awareness of their own scoring procedures. By improving their metacognitive awareness, scorers may be better able to recognize and change their scoring strategies when they begin employing less reliable scoring procedures. Another approach would be to identify training methods that would increase scorers' understanding of and agreement about essay characteristics that are not easily defined such as writing style or the communication of ideas.

References

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.

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

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

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

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

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

Wolfe, E.W. & Feltovich, B. (1994, 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. (1995). *A study of expertise in essay scoring.* Unpublished doctoral dissertation, University of California, Berkeley, CA.

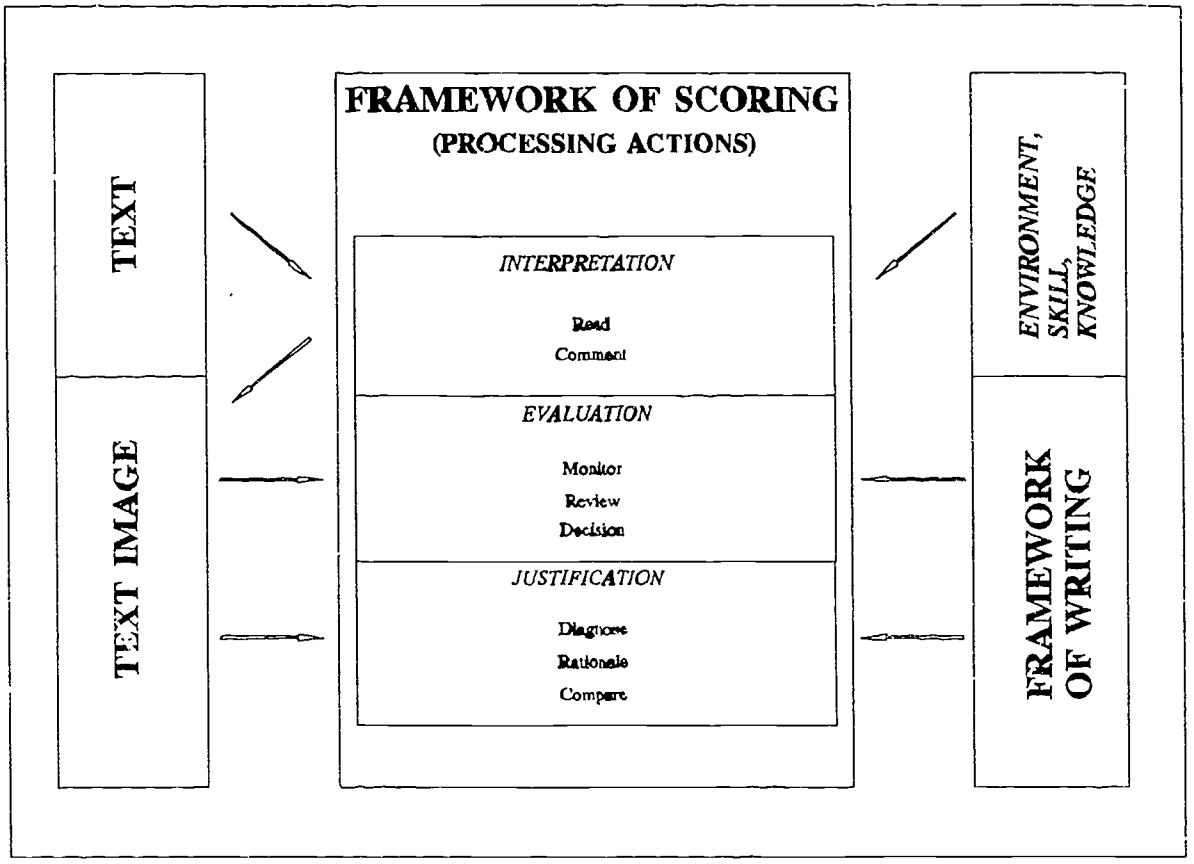

Figure 1: Model of Scorer Cognition

Table 1: Regression Model for Content Focus

Variable	*Standard Coefficient*	*Standard Error*	*t*	*p*
Constant	0.00	0.231	-1.92	.07
Story	0.57	0.002	3.33	.003
Style	0.36	0.002	2.07	.05

NOTE: For the model, F (2,21) = 6.78, p = .005, R^2 = .39

Table 2: Regression Model for Processing Actions

Variable	*Standard Coefficient*	*Standard Error*	*t*	*p*
Constant	0.00	0.167	-0.15	.88
Monitor	0.40	0.002	2.13	.05
Diagnose	0.32	0.005	1.68	.11

NOTE: For the model, $F(2,21) = 3.45$, $p = .05$, $R^2 = .25$