Writing fits well within the realm of outcomes suitable for observation by performance assessments. Studies of the reliability of performance assessments have suggested that interrater reliability can be consistently high. Scoring consistency, however, is only one aspect of quality in decisions based on assessment results. Another is generalizability. Research suggests that if the number of ratings per task could be increased, it may yield an increase in "task" generalizability without a dramatic increase in the actual number of tasks. Multitrait analytic scoring strategies for writing performance assessments may increase "task" generalizability over a single holistic score. Research undertaken by G. Roid (1994) supports the potential usefulness of analytic scores as effective sources for feedback to students and as bases for meaningful discussion on the writing process. Work at the Center for the Study of Evaluation at the University of California, Los Angeles, has expanded on the development of methodology and uses of analytic scoring. Work on narrative-writing-specific scoring rubrics has shown promising evidence of reliability and validity. Training in and use of these rubrics has also increased participating teachers' understanding of the quality components of writing. (Contains 3 figures and 15 references.) (SLD)
A Discussion of Analytic Scoring for Writing Performance Assessments

Kevin D. Crehan
University of Nevada, Las Vegas

The most prevalent response to the call for assessment reform has been to increase the use of more authentic assessments, e.g., performance assessments. Advocates of performance assessments suggest that this form of appraisal can serve to measure important and complex learning outcomes and provide information useful to guide improvement in instruction (Resnick & Resnick, 1989). Perhaps the most complex form of student achievement which we attempt to assess involves composition. Therefore, the task of writing fits well within the realm of outcomes suitable for observation by performance assessments.

Among the problems associated with using performance assessments to measure important learning outcomes are objectivity of ratings and generalizability (reliability) of scores across raters and tasks. A review by Linn (1993) summarized evidence of acceptable generalizability across raters given well-defined scoring rubrics, intensive

rater training, and monitoring during rating. Additionally, the California Assessment Program has established an inter-rater reliability of .90 for their writing assessment by using procedures which include providing sample anchor papers for each rater and recirculating previously scored papers to check on stability (U.S. Congress, Office of Technology Assessment, 1992). Shavelson, Baxter, and Pine (1992) observed the reliability and validity of performance assessments in the 5th and 6th grade science curriculum. They asked the question: How large a sample of observers is needed to produce reliable measurement? Their results found inter-rater reliability to be consistently high in evaluating student performance on complex tasks, high enough to conclude that a single rater provides a reliable score.

While these observations offer promise for the utility of performance assessments, scoring consistency is only one aspect of quality in decision situations based on assessment results. Linn and Burton (1994) suggest that for pass-fail decisions involving individual students, acceptable generalizability across tasks is attained only when a large number of tasks are used, perhaps as many as ten. If the content area is being assessed in writing, such a large number of writing tasks on an occasion might require an unreasonable expenditure of instructional time devoted to assessment to say nothing of the administration and scoring costs. However, if the number of ratings per task could be increased, it may yield an increase in "task" generalizability without a dramatic increase in the actual number of tasks. Multitrait analytic scoring strategies for writing performance assessments may increase "task" generalizability over a single holistic score.

Much of the research on the psychometric characteristics of writing performance
assessments uses single score "holistic" ratings. In writing assessment this single holistic score designed to estimate the wholeness in quality of the writing product. There is agreement (e.g., Huot, 1990) that writing is a multifaceted performance and as such involves attainment on a number of mental traits, e.g., vocabulary, language mechanics (see Figure 1), on which individual differences exist. Additionally, there are different types of writing, e.g., narrative, expository (see Figure 2). Given that writing performance involves a number of traits on which individuals differ, analytic scoring of writing products is recommended by some researchers (see Figure 3) (Roid, 1994; Huot, 1990; Marsh & Ireland, 1987; Novak, Herman, & Gearhart, 1996).

Roid (1994) used cluster analyses to explore the empirical validity of the analytic traits presented in Figure 1. Results of these analyses demonstrated that, while forty percent of the responses had flat trait patterns (either all high or low), a number of distinct patterns among the six traits were evidenced. For example, thirteen percent of the patterns were very close to average on five of the traits but either high or low on conventions. Ten percent of the patterns showed high or low voice, with other scores near average. An additional thirteen percent were either high or low on ideas, organization, and voice but close to average on word choice, sentence fluency, and conventions. This suggests evidence of a creative or stylistic component among the six traits. This evidence supports the potential usefulness of analytic scores as effective sources for feedback to students and as bases for meaningful discussion on the writing process.

Work at the Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, at UCLA (e.g., Wolf & Gearhart, 1993a;
1993b) has expanded on the development of methodology and uses of analytic scoring. Work on narrative-writing-specific scoring rubrics has shown promising evidence of reliability and validity (Gearhart, Herman, Novak, Wolf, & Abedi, 1994; Gearhart, Herman, & Novak, 1996). Additionally, training and use of these rubrics has benefited instruction by increasing participant teachers' understanding of the quality components of writing (Gearhart & Wolf, 1994; Gearhart et al., 1994, Wolf & Gearhart, 1995).
References


**Figure 1**

Definitions of Analytic Traits (Spandel & Stiggins, 1994)

<table>
<thead>
<tr>
<th>Trait</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideas</td>
<td>The heart of the message, the content of piece, the main theme, together with all the details that enrich and develop that theme. Ideas are strong when the message is clear and enlivened with interesting and important details.</td>
</tr>
<tr>
<td>Organization</td>
<td>The internal structure of a piece of writing, the thread of central meaning, the pattern that holds everything together. Organization is strong when the piece begins meaningfully, proceeds logically, and creates a sense of anticipation that is ultimately systematically fulfilled.</td>
</tr>
<tr>
<td>Voice</td>
<td>The writer coming through the words, his or her wit and feeling, the sense that a real person is speaking to us and cares about the message. Good writers impart a personal tone and flavor to the piece that is unmistakably his or her’s alone.</td>
</tr>
<tr>
<td>Word Choice</td>
<td>The use of rich, colorful, precise language that communicates not just in a functional way but in a way that moves and enlightens the reader. Strong word choice may depend more on the skill of using words precisely than on an exceptional vocabulary.</td>
</tr>
<tr>
<td>Sentence Fluency</td>
<td>The rhythm and flow of the language, the sound of word patterns, the way in which the writing plays to the ear - not just to the eye. With good fluency, sentences vary in length and style, and they are so well-crafted that reading aloud is a pleasure.</td>
</tr>
<tr>
<td>Conventions</td>
<td>The mechanical correctness of the piece - spelling, grammar, usage, paragraphing, capitalization, and punctuation. Writing that is strong in convention has been well proofread and edited.</td>
</tr>
</tbody>
</table>
**Figure 2**

Modes of Writing (Roid, 1994)

<table>
<thead>
<tr>
<th>Mode</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive</td>
<td>Describes an object, place, or person, enabling the reader to visualize what is being described and to feel that he or she is very much part of the writer's experience. Writer's purpose is to create a strong and vivid image of impression in the reader's mind.</td>
</tr>
<tr>
<td>Persuasive</td>
<td>Attempts to convince the reader that a point of view is valid or persuade the reader to take a specific action. Writer's purpose is to persuade the reader.</td>
</tr>
<tr>
<td>Expository</td>
<td>Gives information, explains something, clarifies a process, or defines a concept. Writer's purpose is to inform, clarify, explain, define, or instruct.</td>
</tr>
<tr>
<td>Narrative</td>
<td>Recounts a personal experience or tells a story based on a real event. Writer's purpose is to recount an experience or tell a story in a concise and focused way to create some central theme or impression in the reader's mind.</td>
</tr>
<tr>
<td>Imaginative</td>
<td>Tells a story based on the writer's imagination. The story is basically fictional, but the writer may use his or her experience and knowledge of people or situations to bring a special flair or flavor to the writing. Writer's purpose is to entertain the reader or write for the author's own pleasure.</td>
</tr>
</tbody>
</table>
Advantages and Limitations of Multifaceted Analytic Scoring
(Gearhart & Wolf, 1994; Gearhart, Wolf, Burkey, & Whittaker, 1994; Spandel & Stiggins, 1990; Wolf & Gearhart, 1995)

Advantages:

1. Developing the analytic scoring rules forces judgements on what is valued in writing and the product provides an operational definition for the quality characteristics of writing.

2. Allows more systematic and detailed feedback to students on the strengths and weaknesses of their writing.

3. Provides more diagnostic information that teachers may use to guide their instruction and student practice.

4. Benefits the teachers who are trained in the rating method and subsequently perform the ratings. These teachers can use what they learn to improve their writing instruction and feedback to students.

5. Ratings on multiple facets of the domain of writing skills allows improved generalizability over a single holistic score.

Limitations:

1. Analytic scoring can be very expensive and time consuming if not well managed.

2. The analytic rating task is not for everybody. The rating task is initially difficult and beginning raters may experience frustration.
I. DOCUMENT IDENTIFICATION:

Title: A Discussion of Analytic Scoring for Writing Performance Assessments

Author(s): Kevin D. Crehan

Corporate Source: University of Nevada, Las Vegas

Publication Date: October, 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

[ ] Sample sticker to be affixed to document

Check here

“PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)”

Level 1

[ ] Sample sticker to be affixed to document

Sample

“PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)”

Level 2

or here

“PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)”

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Kevin D. Crehan Position: Associate Professor

Printed Name: Kevin D. Crehan Organization: University of Nevada, Las Vegas

Address: College of Education Telephone Number: (702) 895-4303

Mail stop 3003 Date: 10-29-97

Las Vegas, NV 89154