UC Davis
Journal of Writing Assessment

Title
The Effect of Scoring Order on the Independence of Holistic and Analytic Scores

Permalink
https://escholarship.org/uc/item/4xb1c9gj

Journal
Journal of Writing Assessment, 4(1)

Authors
Singer, Nancy Robb
LeMahieu, Paul

Publication Date
2011

Copyright Information
This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, availalbe at https://creativecommons.org/licenses/by-nc-nd/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

The Effect of Scoring Order on the Independence of Holistic and Analytic Scores

by Nancy Robb Singer and Paul LeMahieu

Abstract

Conventional wisdom and common practice suggest that to preserve the independence of holistic judgments, they should precede analytic scoring. However, little is known about the effects of scoring order on the scores obtained or if true holistic scoring is even possible from the mind of a scorer who has already been trained to and will be asked to provide analytic scores as well. This research explores the matter of independence of scores and the effects of scoring order upon those judgments. Our analysis shows statistically significant differences in mean scores under the two conditions (holistic scoring preceding analytic and the reverse), with the holistic scores more nearly replicating "pure" holistic scoring only when it precedes the analytic. This research affirms that when readers will be asked to score both ways, holistic scoring should precede analytic scoring. It also suggests interesting insights into the cognitive processes engaged by scorers as they score holistically and analytically.

When assessing student writing, there are often legitimate reasons for seeking both a holistic score as well as a set of analytic scores. Holistic scoring refers to assessments in which a single summary judgment of quality is rendered, albeit often guided by a conceptual framework that articulates essential dimensions upon which quality is to be defined (Huot, 1988; White, 1984). In analytic scoring, writing is described in terms of essential attributes and individual judgments recorded for several constituent attributes. Holistic scoring is thought by some to enable a more complete and appropriate depiction of a phenomenon as complex as writing, which they feel can never be adequately deconstructed into several parts (White, 1994). Moreover, singular summary judgments may fulfill certain assessment purposes perfectly well, while being quicker and, therefore, more economical (Spandel & Stiggens, 1980). Analytic scoring, on the other hand, may be considered more informative and therefore potentially relevant to instructional and programmatic decision making, though it is more time consuming and therefore generally more expensive to conduct (Cooper & Odell, 1977; Rabianski, 1979).

As a practical matter, instructional interests coupled with economic or logistic concerns often dictate that both types of scores be obtained in a single common scoring session. A recent survey of state writing assessment systems in the United States (National Writing Project [NWP], 2008b) shows that 46 of the 50 states have direct writing assessments. Of these, 32 (67%) assign holistic scores and 21 (44%) assign analytic scores. Seven states assign both types of scores in their writing assessment systems. When both kinds of scores are desired, questions reasonably arise regarding the independence of the scores obtained. Because holistic and analytic scores have very different conceptual meanings, it is essential to the validity of each type of score evidence that it not be unduly influenced by the other. For example, to those deeply concerned about holistic assessment, such scores are not merely the aggregate (be it a sum or average) of some set of analytic scores. Rather, a holistic score is conceived of as something qualitatively different than the mere aggregate of many parts. The written product may weave those parts together in a way that creates a greater (or conceivably lesser) whole.

To preserve the greatest degree of independence of holistic judgments, rational analysis would suggest that the "whole" of the writing should be judged first before scoring component parts through analytic scoring. However, little is known about the effects of scoring order on the scores obtained or if true holistic scoring is even possible from the mind of a scorer who has already been trained to and will be asked to provide analytic scores at the same time. This research explores the matter of independence of judgments and the effects of scoring order upon those judgments.

Background

While there is much research and discussion on the merits and pitfalls of both analytic and holistic scoring (e.g., Weigle, 2002; White, 1994; Wolcott & Legg, 1998), and on the reliability and validity of these methods of assessment (e.g., Camp, 1993; Yancey, 1999), there is far less research that considers the cognitive processes of readers and the effects on resulting scores. A sample of research using protocol analysis that explores the complexity of scorers' decision making includes Pula and Huot (1993) who examined how scorers' personal backgrounds, education, and professional experience influenced their reading of student writing. Wolfe's (1997) research used protocol analysis to find that raters who agreed at a high rate in large-scale scoring sessions read more quickly and narrowly than other raters, and these "effective" raters focused in their talk-alouds almost exclusively on rubric-related features of student writing. And, DeRemer (1998) looked at scorers' processes as they defined assessment tasks. In our research, we extend these prior studies to consider how the scoring process may be impacted when readers are asked to score both holistically and analytically in the same setting. Additional research (e.g. Freedman, 1981; Veal and Hudson ,1983; Hunter, Jones, and Randhawa, 1996) suggests that because there is a demonstrable correlation between holistic and analytic scores, the cheaper and less time-consuming method of holistic scoring is sufficient.

The data analyzed in this research was the result of both holistic and analytic scoring. The National Writing Project sponsored the scoring as part of a research initiative examining the impact of various aspects of its programming. In doing so, it conducted both kinds of scoring as it sought statements about differences in overall performance (per holistic scoring), as well as a close analysis of specific areas of impact on writing performance as examined by analytic scoring.

National Writing Project

Founded in 1973, the National Writing Project (NWP) is the premiere professional network for teachers concerned about writing, the teaching of writing, and the use of writing in teaching. NWP emphasizes core principles of effective instruction while attending to local needs, reform priorities, and school conditions in over 200 university-based writing project sites in all 50 states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands and a number of foreign nations.

The NWP model begins with an annual four- to six-week invitational summer institute at each local writing project site, attended by experienced teachers of all grade levels and disciplines, including postsecondary colleges and universities. These teachers prepare for leadership roles by studying research and reflecting upon its application to their own practice (Lieberman and Wood, 2003); engaging in a scholarship of practice that includes the sharing (publication), examination (peer review), and adaptation (adoption) of their most effective practices (Blau, 2003; Shulman, 1993; Stock, 2001) and improving their knowledge of the learning and teaching of writing by becoming writers themselves (Whitney, 2006). After the institutes, these teachers become teacher-consultants of the National Writing Project and join a large and growing reservoir of local capacity, conducting project-sponsored professional development programs in neighboring schools and districts (Gray, 2000).

A central feature of the NWP model is that the specific design of professional development programs varies according to local needs, reform priorities, school conditions, and research contexts. Each writing project site emphasizes common core principles of effective instruction, with the design and delivery of specific activities and services negotiated with local education authorities. Beginning in 2003, NWP began supporting research studies designed to address evaluative questions examining the impact of NWP activity through its Local Sites Research Initiative (LSRI). The National Writing Project's goal for the LSRI was to develop a growing body of research that examined local professional development programs based on these core principles and that illuminated teacher practices and student achievement in writing across a range of grade levels, schools, and local contexts. Readers wishing to view the results of these LSRI studies will find them at http://www.nwp.org/cs/public/print/programs/lsri (Buchanan et al., 2006).

Scoring Conference

To help local writing project sites evaluate their programmatic efforts and their impact on student achievement, NWP conducted a large-scale writing assessment in the summer of 2005. Local site research teams administered common writing prompts to their target groups (elementary, middle, or high school) as well as carefully matched comparison groups in a pre-post measure design. Writing project sites then submitted over 7,500 writing samples to NWP, which convened the scoring conference to ensure the technical rigor, independence, and credibility of the writing scores employed in the sites' research studies. The students' written product was integrated across sites and scored without any information identifying the sample as to individual, site, group (program or comparison), or time of administration. A common evaluative framework and scoring system (NWP, 2005, 2006, 2007, 2008a) was employed for all scoring, with differentiated standards established and maintained through the anchor papers, scoring commentaries, and related training and calibration materials used for each grade span (elementary, middle, and high school). Scorers were all considered to be "expert" in terms of personal background, professional training, and work experience (Pula & Huot, 1993; Wolfe, 1997). To train for the scoring session, raters were apprised of procedures and logistics for scoring the student writing, reviewed and discussed the rubrics and anchor papers, and trained together and individually by scoring sample student writing. The training typically lasted approximately six hours until scorers were calibrated to the criterion level of agreement in scoring. Subsequently, following every major break in scoring (e.g., overnight, after meals, etc.) all scorers were recalibrated once again to the criterion level of performance before operational scoring resumed.

To assist writing project sites in assessing their students' writing achievement and to adequately address the complexity of the sites' research questions and the students' writing performance, the national scoring conference sought both holistic and analytic scores. It is important to note that these local sites' initiatives were not part of a high-stakes testing program; no judgment about individual student performance was rendered. Rather, the resulting data was used in program evaluation and in efforts to plan for future instruction and program development.

The order of the assessment of student writing was guided by a belief that holistic scores should be assigned before the analytic ones to maximize their independence. This research uses the existing dataset from the Local Site Research Initiative to examine that conventional wisdom by varying the conditions under which the papers were scored. A set of randomly selected papers was scored under varied conditions (i.e., holistic then analytic, analytic then holistic, "pure" holistic, and "pure" analytic) to assess the potential impact scoring order on the assessment of student writing.

Guiding Questions

Analysis of data allowed us to explore these questions:

1. Does the condition under which scorers read (holistic then analytic, analytic then holistic, "pure" holistic, "pure" analytic) affect the final score of the paper?
2. In terms of reliability of scoring and differences in scores assigned, what impacts can be associated with the various conditions (order of scoring) examined?
3. What do related qualitative data--think-alouds, questionnaires, interviews--suggest about the cognitive and decision making processes of readers engaged in scoring?

Methods

Sample.

This research examined 257 papers selected from the more than 7,500 scored at the NWP national scoring conference in the summer of 2005. This sample size was determined by conducting a power analysis, which revealed that acceptably stable results could be obtained in an analysis based upon anything over 250. Using the variability in previous scoring as a best estimate of the inherent variance in scores and employing a desired level of confidence of $\alpha \leq .05$, this sample size was projected to be sufficiently sensitive to yield an efficient test. The observed results in this research bore this out.

The papers were all from the middle school grades (grades 6-8) because it was the grade level within which the highest levels of reliability were achieved for the scoring. It was also thought to be the level for which the experience in this instance would most profitably inform the other grade levels. A stratified random sampling design was employed with writing performance as the stratification factor. Thus, the resulting random sample represented the full range of writing performance so as to ensure that level of performance did not confound the results observed.

Data Collection.

This study examined both quantitative data (i.e., the scores awarded by trained assessors of writing under various conditions) as well as qualitative data (such as those derived primarily from observations of training and scoring sessions as well as think-alouds taken from scorers as they engaged in evaluating student writing).

The assessment system used in the research was one developed by the NWP through extensive modification of the *6+1 Trait Model of Writing* (Culham, 2003; Bellamy, 2005). The assessment system was developed in the first instance to support a program of research evaluating NWP programmatic impact (see, for example, Buchanan et al., 2006 and NWP, 2008c). The broad applicability of the framework was a major appeal of the *6+1 Trait Model*, as was its widespread use in schools and school districts. The *6+1 Trait Model* provides an analytic rubric attending to six specific attributes as well as the overall character of students' writing (Buchanan et al., 2006). While sufficiently comprehensive, certain attributes of the model rendered it inappropriate for the research purposes of the LSRI, and a number of modifications were made to adapt it for use in research studies. The following modifications were made in the rubric for both the holistic as well as the analytic versions:

- The scale of the rubric was extended from four to six points in order to ensure sufficient discrimination and therefore to allow increased sensitivity to any changes in student performance over time or differences between groups (and most certainly the interaction between the two as was the focus of the evaluative research).
- The language defining the traits was clarified to enhance the reliability of evaluative judgments.
- The focus of the evaluative judgments was modified to examine exclusively the student writing (where, on occasion, the original rubric included references to the reader's reactions or to perceptions about the writer's engagement interest in the topic as the basis for judgment).

The resulting measure examined the following attributes. While extensively modified since its use in this research, operational definitions of attributes can be found in Appendix A.

- *Ideas/Content Development*: establishing purpose, selecting and integrating ideas, including details to support, develop, or illustrate ideas;
- *Organization*: creating an opening and closing, maintaining focus, ordering and relating events, ideas, and details to provide coherence and unity in the writing;
- *Voice*: communicating in an engaging and expressive manner, revealing the writer's stance toward the subject;
- *Sentence Fluency*: constructing sentences to convey meaning, controlling syntax, and creating variety in sentence length and type;
- *Word Choice*: choosing words and expressions for appropriateness, precision, and variety;
- *Conventions*: controlling grammar, punctuation, spelling, capitalization, and paragraphing.

In addition to scores in each of these areas, each writing sample received an overall holistic score, one conceived not as an aggregate of these component parts but as an independent (in the psychometric sense of being independent of the other scores assigned), overall summary judgment.

All of the papers included in this study were double-scored to ensure the quality of the scores included in the subsequent analyses. Table 1 presents the reliabilities of scores for the middle school scoring under all three conditions as well as across all scorings. In this instance, reliability was defined as interrater agreement, with agreement defined as both identical and immediately adjacent scores. For the middle school sample, reliabilities ranged from 91 to 100% under the "conventional" conditions of holistic scoring first then analytic (with an aggregate of 95%), and from 91 to 98% (with an aggregate of 95%) under the condition of "reverse" scoring (analytic first then holistic). The reliabilities of the "pure" scorings were similarly high, ranging from 87 to 95% for the pure analytic scoring (with an aggregate of 91%). The pure holistic scoring was conducted with a reliability of 94%. These levels of reliability are as high as is typically observed, and are easily adequate to support the research purposes pursued here.

Table 1: ***Reliability of Scoring under Various Conditions***

Level	Number of contrasts	Total % agree	Holistic % agree	Ideas % agree	Organization % agree	Voice % agree	Sentence Fluency % agree	Word Choice % agree	Conventions % agree
Holistic then Analytic	1,799	95%	96%	95%	93%	91%	93%	96%	100%
Analytic then Holistic	1799	95%	95%	95%	91%	91%	91%	94%	98%
Pure Analytic	1542	91%		90%	89%	87%	93%	95%	90%
Pure Holisitic	257	94%	94%						

N= 257 with all papers double scored.

The qualitative data for this study included independent observer's field notes taken during scorer training sessions, follow-up open-ended questionnaires distributed after the scoring, and think-alouds conducted with eight randomly selected participants as they scored student writing. Think-alouds have been widely used to explore the cognitive task of scoring (e.g. Newell & Simon, 1972; Pula & Huot, 1993; Wolfe, 1997). While we acknowledge the criticism and constraints of using think-alouds (e.g. Ericsson & Simon, 1984; Smagorinsky, 1989; Steinberg, 1986) and therefore temper our insights drawn from them, we believe that the use of think-alouds in this study gave us a glimpse of the heavy cognitive demands placed on scorers as they attempted the complex task of scoring both holistically and analytically. Because the sample size for the think-alouds was small, these qualitative data were never intended to be a direct unit of analyses--statistical or otherwise. Rather, the purpose of these data was to inform inferences made based upon the quantitative data and to suggest reasonable interpretations of those results.

Scorers met individually with the researcher during the second day of scoring. The timing both ensured that readers were well-versed in and comfortable with the scoring guides and that the think-alouds did not interfere too greatly with the primary goal of completing the necessary scoring. During these taped and transcribed think-alouds, scorers were read verbal instructions that asked them to "stop each time [their] internal voice registered an opinion, asked a question, or made a judgment." They responded to the prompt soliciting their thoughts and reactions to the writing in situ in as close to real time as is possible. Following the think-aloud, scorers were asked a series of four scripted interview questions as well as individual follow-up probing questions. (See Appendix B.)

Procedures

Since the purpose of this research was to examine the effect of scoring order on the scores assigned to student work, each paper was scored under four separate conditions:

1. *Holistic then analytic (H→A)*: These scorers first assigned a holistic score and then a set of analytic scores. This is the condition most widely used when both holistic and analytic scores are to be assigned by a single scorer. Common practice is predicated on the belief that it is easier to assess a "whole" before one assesses its analytic "parts" and that the likelihood of achieving independently assigned scores in the psychometric sense of local independence--wherein one score is not influenced by others--is high under this condition.
2. *Analytic then holistic (A-->H)*: A second group of scorers first assigned analytic scores and then a holistic score. These scorers were trained to score only under this one condition.
3. *"Pure" holistic (P[H])*: A third independent group of readers assigned only a holistic score. They were trained only to use this one system and never scored under any other condition.
4. *"Pure" analytic (P[A])*: A fourth independent group of readers assigned only a set of analytic scores. They were trained only to score under the analytic condition and never scored in any other condition.

Thirty-four individuals trained to score under the first two conditions at the middle school level. All 34 of these scorers received the same training. That is to say, they sat next to each other in the same room with facilitators who trained using the same rubrics, the same anchor papers, the same training papers and scoring commentary, as well as the same procedures for all scorers. After the training to criterion levels of reliability, operational scoring proceeded as usual for the majority of scorers (under Condition 1 [H-->A]). However, six of the teachers were randomly assigned to score under Condition 2(A-->H). To the extent possible, these procedures ensured that training effects were minimized.

The raters who scored under Conditions 1 and 2 were highly experienced (mean years in teaching = 17.8); all of them were trained teacher-consultants with the National Writing Project; and a substantial number, 13 of 34 (or 38%), had prior formal experience with writing assessments, having been trained for and participated in scoring within state or local writing assessment systems.

Three months later in a different location and with 13 different scorers, the same student papers were subsequently rescored to provide scoring data under the third and fourth conditions (P[H] and P[A]). These scorers received similar training. They participated in the same basic training, provided by the same trainers, and using the same anchor papers and training materials. However, depending on the scoring condition to which they had been randomly assigned, teachers were trained only to use either the holistic or the analytic rubrics.

The independent sets of teachers who later scored under the two pure scoring conditions (P[A] and P[H]) were similarly experienced and accomplished. Again, these scorers were teacher-consultants with the National Writing Project; they had a mean number of years in teaching of 16.5; and 31% (4 of 13) had trained and participated in formal assessment systems at their state or local levels.

This distribution of scorers (28 randomly assigned to Condition 1; six to Condition 2; five to Condition 3; and eight to Condition 4) was determined by the different scoring loads under each condition. The goal was to complete all the scoring within a similar time frame so as to minimize fatigue effects, yet the amount of scoring is different under each of the conditions. Hence, the different numbers of scorers adjusts for these differences. This explains the balancing of the numbers of scorers for all conditions except Condition 1, in which such a large number of scorers were employed because the scoring of writing samples under the traditional condition included many more samples from the whole of the evaluation research--the primary purpose for the scoring within which this assessment research was embedded. What is material in the preparation and assignment of scorers is that the identical training given, the random assignment of scorers to scoring conditions, and the balancing of the numbers assigned to each writing sample all was intended to minimize any possible effects of scorer characteristics, scorer training, or scorer fatigue as possible challenges to observed results.

For each of the four conditions, each paper was scored twice with discrepancies (scores that differed by more than one scale point) being adjudicated by a third reader. The adjudication score identifies one of the first two scores as an outlier and replaces it to arrive at the two operational scores. The resulting final score used in subsequent analyses is the average of the two operational scores assigned to each paper. All scores were then entered into a database in which the individual papers became the unit of analysis with sets of separate scores--confirming to the four scoring conditions--assigned to each paper. Those scores then became the dataset used in analysis.

Data Analyses

Quantitative data were analyzed using a repeated measures analysis of variance. The units of analyses were the 257 writing samples and the within unit factor was the condition or method of assessment (H-->A, A-->H, P[H] or P[A]). Following on the analyses of variance, a Scheff&#65533; post hoc comparison was examined for each analysis of variance for which significant differences were observed in the omnibus test. These contrasts were examined to identify which condition represented the locus of any observed significant differences.

Results

Table 2 provides the results of a repeated measure analysis of variance on these data. Presented there are the seven scores and the relevant means for each of the four conditions of scoring (H -->A, A-->H, P[H], and P[A]). To the far right is the F statistic from the repeated analysis of variance and its attendant level of significance. The omnibus test of differences between the means across conditions is highly significant (alpha < .0005) for every measure with the exception of conventions, the only attribute for which the differences were not significant.

The reader will see in Table 2 that when analytic scoring preceded holistic scoring (A-->H) the mean scores were different to such a degree as to be statistically significant. This was true for all scores with the only exception being those for conventions. Differences were not statistically significant for the conventions trait and this was similar across all conditions of scoring.

Table 2: ***Repeated Measures Analyses on Method of Scoring (N = 257) 1***

	Mean Score	by Condition			
Measure	A → H	H → A	Pure	*F*	P(F)
Holistic	3.00	2.77	2.79	18.31	< .0005
Ideas	3.22	2.89	2.88	39.01	< .0005
Organization	2.99	2.70	2.61	41.12	< .0005
Voice	3.38	3.08	3.04	30.23	< .0005
Sent Fluency	3.09	2.82	2.65	46.39	< .0005
Word Choice	3.06	2.80	2.75	31.02	< .0005
Conventions	2.93	2.90	2.90	00.40	< .674

1 Scheff&#65533; post hoc comparisons were conducted to determine the locus of differences where the omnibus test revealed significant differences. Those values identified by underlining the mean score for the conditions that were significantly different from the other two.

Finally, the data obtained for this research allows for the construction of composite scores, allowing for comparisons among the scores derived from the aggregation of pure analytic scoring and holistic scores. The composite score is the mean of the component parts (traits), which can in turn be compared to a holistic score (derived under pure conditions as well as each of the combined holistic-analytic scoring conditions). For each of the three conditions, Table 3 provides both a mean composite score and a mean holistic score. To the far right is the relevant F statistic and its level of significance. Under all three conditions the composite score is higher than the holistic score and that difference is statistically significant in three of the four conditions. In fact, it is significantly higher in the conditions where both forms of assessment--holistic and analytic--were pursued. It is in the "pure" assessment where the difference between the pure holistic score and the composite score derived from pure analytic scores is not statistically significant from one another.

Table 3: ***Repeated Measures Analyses on Holistic and Composite Scores (N = 257)***

	Mean	Score		
Condition	Composite	Holistic	F	P(F)
A-->H	3.11	3.00	50.92	< .0005
H-->A	2.86	2.77	28.45	< .0005
Pure	2.81	2.79	.137	.712

Discussion

One of the primary purposes of this research was to examine the conventional wisdom that holistic scoring is best done before analytic scoring when scorers are called on to do both. The Scheff&#65533; post hoc comparisons examining the significant differences among the mean holistic scores by condition locates those differences between the A-->H condition and both H-->A and P(H), while the holistic scores, when scoring H-->A, are not significantly different from the pure holistic scores.

One might legitimately assert that the pure holistic score is the preferred standard against which the holistic scores derived from other conditions should be compared, as it is the best approximation of what holistic scores would be assigned when holistic scoring alone is conducted. The absence of significant differences between scores from the H-->A scoring condition and the pure holistic scoring (coupled with the fact that it is the resultant A-->H score that is significantly higher than the others) bears out what is common practice--not just as the best option, but in fact as the one in which the scores will essentially replicate pure holistic scoring.

Table 2 also reveals that the A-->H scoring condition results in significantly higher scores than either of the other two conditions for all measures except the conventions score. The relatively consistent alignment of the scores within condition and the fact that the significant differences across conditions are located between the A-->H condition and the other two bear some explanation.

It would seem, first and most simply, that there is a tendency for alignment within condition. Scorers tend to assign scores in which the first form of scoring (be it holistic or analytic) establishes the frame within which the subsequent scores are aligned. The think-alouds reveal that scorers were cognizant of the inherent differences between the two modes of scoring and tried hard to stay within the designated scoring form. In her think-aloud, one scorer gave this description:

> Just the holistic six-point scale wasn't enough. I had to turn to the more detailed development of those things. There's a part of me that wonders if I'm moving into the analytic realm as soon as I start doing that. But because I'm not trying to weight, I'm hoping that I'm just using this to help me define what these are.

This is intuitively understandable, and it substantiates the importance of doing the holistic scoring first both to protect its independence (to the extent possible) and insofar as it is under this condition that the resultant holistic score most closely resembles the pure holistic score.

What is less apparent is why scores from the A-->H condition are so consistently higher than the others (and yet with the conventions score not so). We believe that when scoring the student writing, these teachers (while trained to the point of exemplary consistency) approach their task as a "satisficing" activity (Simon, 1957). "Satisficing" refers to a circumstance in which decisions must be made with less than perfect or complete information. Under such conditions, expert performers identify and seek that

information (in both kind and amount) that enables a valid and credible decision. Satisficing, we believe, is one that experienced classroom teachers (who made up our scorer pool) commonly use in their daily practice. That is to say, their professional judgment about classroom lessons, student performance, etc., is almost always rendered with less than complete information. Applied to this scoring context, this concept of satisficing suggests that scorers may seek evidence of accomplishment to some level of performance, and that level (referenced against the rubric and the anchors) determines the assigned score. This is a perfectly valid activity as long as the search for satisficing evidence involves attributes of the writing that are demonstrably present. In other words, this satisficing approach can yield valid and critically honest scores as long as it is based upon bona fide relevant evidence and the score assigned goes no higher than such evidence allows.

Nonetheless, scoring approached in this manner could indeed result in a different score than one in which the mental process was one of uncovering evidence of need or shortcoming. A reader who scored in the H -->A condition expressed a tension in wanting her holistic and analytic scores to "match." She said,

> Knowing that I would score [both holistically and analytically] influenced me in that I expected the score to be fairly consistent... I was able to hold one system separate from the other while I read; however, one did make me second guess my score based on the score I had given with the first reading... I had to remind myself that the scores did not have to match.

This satisficing activity results in generally higher scores when doing the analytic scoring first inasmuch as it is a mental process that is repeated a number of times for each of the several attributes. By contrast, the singular summary judgment of the holistic score (even when approached in this satisficing manner) does not provide as much particular opportunity for the detection of such evidence. A reader who scored in the A-->H condition said,

> From the papers I was scoring, I thought I was really too hard and I [thought the papers needed] to come up. [I thought] 'it does have this' or 'it does have that.' Something in my head said, 'If you're leaning more toward the higher end, go with it.'

This tendency to "go higher" seemed to be especially prevalent with "cusp" scores--that is, when scorers were mentally debating between, for example, a 3 or a 4 score on the analytic rubric. One reader said this about score changing: "It tended to be to people who were on the cusp... If I could have given a 3.5, then I would have felt comfortable with that." Under these circumstances, scorers appeared to be more willing to give the student the benefit of the doubt on the analytic rubric and assign a higher score.

Time spent on the analytic and holistic scoring may also shed light on the differences in scores. In all conditions, and as one might expect, readers scored much faster holistically than they did analytically. However, after scoring analytically, participants in the A-->H condition arrived at the holistic score with extraordinary speed. Although time was not measured or tracked in this study, in one think-aloud that may well be exemplary, the transcript revealed two single-spaced pages of analytic decision making and only one sentence for the holistic judgment. The scorer said, "I think I do take more time to just question myself when I do it analytically first." When it was time for scorers to render a holistic score, there was little mental tabulation of the paper's strengths and weaknesses. Another reader scoring in the A-->H condition said,

> It was easy to forget about the holistic score when I graded for the analytic scores first. The analytic scores required me to really focus since I was looking for such specific things. Adding on the holistic score after felt very natural, like icing on the cake.

When scoring in the A→H condition, think-alouds revealed that scorers often accreted evidence of performance along the way and thus the holistic score was assigned quickly.

The fact of no differences for the conventions score among or between the conditions may implicitly support the satisficing interpretation as well. The conventions score is the one that most involves error detection, often resulting in a quantitative basis for the judgment. A threshold degree of performance on conventions compels the assignment of the corresponding score regardless of whether this analytic attribute is scored first or second. The less interpretive stance with regards to the relevant evidence leaves little room for a satisficing approach to the scoring. It is for an understandable reason that this attribute alone results in similar scores regardless of the order of scoring. Transcripts of the think-alouds also suggest that conventions is a trait on which scorers did little mental deliberation. They quickly arrived at a score and rarely changed their minds.

Conclusions and Further Research

The primary purpose of this research was to examine the effects of scoring order when readers are asked to use both holistic and analytic systems of scoring. To determine if scoring order matters, we had scorers rate a common set of papers under one of three

conditions: holistic then analytic; analytic then holistic; and pure holistic or pure analytic. Our data show that when analytic scoring preceded holistic scoring (A-->H) the mean scores (with the exception of conventions) were higher to a statistically significant degree. The reverse condition, holistic preceding analytic (H-->A), resulted in mean scores that were statistically similar to the pure scoring.

This research affirms the common wisdom that readers will more validly score holistically when they can first assess a piece "as a whole," that is, before it is scored for its analytic, component parts. This research also suggests that holistic scores thusly obtained can be construed to be equivalent to independently obtained scores in that they most closely replicate the independent pure holistic scores. Those in charge of designing and conducing large-scale writing assessments that call for both a holistic and analytic score should be aware that the most prudent order of these tasks--as demonstrated in the statistics provided above--is to have readers first score holistically, then score analytically.

While affirming the advisability of the common practice that H-->A scoring is preferred when both are done at the same time, for the first time this research quantifies the difference between scores derived under the various conditions. In doing so, it suggests that the distinction is by no means a trivial one. Scorers' adherence to the suggested order of scoring (verified repeatedly in the think-alouds) coupled with the obvious effects attached to those orderings are worthy of further study. It is possible that there may be effects observed within analytic scores, contingent upon the order in which the analytic elements are examined. It is possibly not a conscious decision that the analytic elements are ordered with ideas, organization, and voice at the beginning and conventions at the end (although certainly neither is it by chance). These results, while not illuminating the issue of order of individual traits directly, do strongly support the importance of the question and also suggest the need for further research.

References

Bellamy, P. C. (Ed.). (2005). *Seeing with new eyes*. Portland, OR: Northwest Regional Educational Laboratory.

Blau, S. (2003). *The literature workshop: Teaching texts and their readers*. Portsmouth, NH: Boynton/Cook.

Buchanan, J., Eidman-Aadahl, E., Friedrich, L., LeMahieu, P., & Sterling, R. (2006). *National Writing Project Local Site Research Initiative report, cohort II, 2004-2005*. Berkeley, CA: National Writing Project.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M. M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. (pp. 45-78). Cresskill, NJ: Hampton Press.

Cooper, C. R., & L. Odell. (1977). *Evaluating writing: Describing, measuring, judging*. Urbana, IL: National Council of Teachers of English.

Culham, R. (2003). *6 + 1 Traits of writing: The complete guide*. New York: Scholastic, Inc.

DeRemer, M.L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, 5(1), 7-29.

Ericsson, K.A. & Simon, H.A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge: MIT Press.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology*, 71, 328-381.

---. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English*, 13(3), 245-255.

Gray, J. (2000). *Teachers at the center*. Berkeley, CA: National Writing Project.

Hunter, D. M., Jones, R., & Randhawa, B.S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation*, 11(2), 61-85.

Huot, B. (1988). *The validity of holistic scoring: A comparison of the talk aloud protocols of expert and novice holistic raters*. (Doctoral dissertation, Indiana University of Pennsylvania).

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

Lieberman, A., & D. Wood. (2003). *Inside the National Writing Project: Connecting network learning and classroom teaching*. New York: Teachers College Press.

National Writing Project. (2005, 2006, 2007, 2008a). The Analytic Writing Continuum: A comprehensive writing assessment system. University of California, Berkeley; Berkeley, CA: National Writing Project.

National Writing Project. (2008b). An analysis of scoring systems employed in state writing assessment programs throughout the United States. University of California, Berkeley; Berkeley, CA: National Writing Project.

National Writing Project. (2008c). Writing project professional development for teachers yields gains in student achievement: Research brief. University of California, Berkeley; Berkeley, CA: National Writing Project.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

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

Rabianski, N. (1979). *An exploratory study of individual differences in the use of freewriting and the use of tagmemic procedures*. (Doctoral dissertation, State University of New York at Buffalo).

Shulman, L. (1993). Teaching as community property: Putting an end to pedagogical solitude. *Change*, 25, 6-7.

Simon, H. (1957). *Models of man*. London: John Wiley & Sons.

Smagorinsky, P. (1989). The reliability and validity of protocol analysis. *Written Communication*, 6 (4), 463-479.

Spandel, V. & Stiggens, R.J. (1980). *Direct measures of writing skill: Issues and applications*. Portland, OR: Northwest Regional Educational Development Laboratory.

Steinberg, E. R. (1986). Protocols, retrospective reports, and the stream of consciousness. *College English*, 48 (7), 697-712.

Stock, P. (2001). Toward a theory of genre in teacher research: Contributions from a reflective practitioner. *English Education*, 33(2), 100-114.

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English*, 17 (3), 290-296.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

White, E. M. (1984). Holisticism. *College Composition and Communication*, 35(4), 400-409.

White, E. M. (1994). *Teaching and asssessing writing: Recent advances in understanding, evaluating, and improving student performance*. San Francisco: Jossey-Bass.

Wolcott, W., & Legg, S. M. (1998). *An overview of writing assessment: Theory, research and practice*. Urbana, IL: National Council of Teachers of English.

Wolfe, E. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, 4(1), 83-106.

Whitney, A. (2006). The transformative power of writing: Teachers writing in a National Writing Project summer institute. (Doctoral dissertation, University of California, Santa Barbara).

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, 50, 483-503.

Appendix A
Descriptions of Scoring Categories*

Content
The content category describes how effectively the writer establishes a purpose, selects and integrates ideas related to content (i.e., information, events, emotions, opinions, and perspectives) and includes details (i.e., evidence, anecdotes, examples, descriptions, and characteristics) to support, develop, and/or illustrate ideas in a unified manner that considers the audience.

Organization
The organization category describes how effectively the writer creates an opening and provides closure; establishes and maintains a focus; orders and arranges events, ideas, and/or details within each paragraph and within the work as a whole; establishes relationships between events, ideas, and/or details within each paragraph and within the work as a whole, promoting coherence and unity.

Voice
The voice category describes how effectively the writer communicates in a manner that is expressive and engaging, thereby revealing the writer's stance toward the subject. Voice is evident when a writer shows a sense of his/her personality through the writing.

Sentence Structure
The Sentence Structure category describes how effectively the writer constructs sentences to convey meaning. It includes the writer's ability to control syntax (i.e., the use and arrangement of words to form a sentence with the proper use of punctuation and capitalization, the arrangement of sentences within a paragraph) and to create variety in sentence length and type (i.e., simple, compound, and complex sentences). Sentence construction gives the reader a sense of the writer's style.

Word Choice
The Word Choice category describes how effectively the writer chooses words and expressions for appropriateness, precision, and variety. Word choice can create powerful imagery (i.e., it should help the reader picture people, places, and objects, and sense feelings written about by the author). Figurative language (i.e., similes, metaphors, and personification) help create vivid images.

Conventions
The conventions category describes how effectively the writer controls grammar, punctuation, spelling, capitalization, and paragraphing. Conventions affect readability.

* These scoring categories have undergone extensive revision for its current operational version. The categories are presented as they appeared at the time of this research.

Appendix B

Think Aloud Follow-up Questions

1. Which did you find it easier to train for--analytic or holistic? (Probe: Why?)

When you scored, were you asked to assign an analytic or holistic score first?

2. How "comfortable" or "natural" does this scoring arrangement seem to you?
(Probe: What are the strengths/challenges to scoring in each condition?)

3. Because you know you are going to score both ways, do you feel this influenced your reading and/or your scoring?
(Probe: How?)

4. To what extent are you able to hold one system separate from the other while you read?
(Probe: How do you hold them separate?)

Bio
Nancy Robb Singer is associate professor of English and Education at the University of Missouri-St. Louis, where she also directs the Gateway Writing Project. Her research interests include teacher preparation and induction, composition theory and research, educational technology in teacher education and in composition, and writing assessment.

Paul G. LeMahieu is Senior Managing Partner for Design, Development, and Research at the Carnegie Foundation for the Advancement of Teaching and graduate faculty in education at the University of Hawai'i - Mānoa. He has published extensively on issues such as educational assessment and accountability as well as classroom learning and the professional development and policy environments that support it. Prior to that, LeMahieu was Director of Research and Evaluation for the National Writing Project and Superintendent of Education for the State of Hawai'i.

Nancy Robb Singer
Associate Professor, English/Education
University of Missouri-St. Louis

443 Lucas Hall
St. Louis, MO 63121
singerna@umsl.edu
314/516-5517

Paul LeMahieu
Carnegie Foundation for the Advancement of Teaching
51 Vista Lane
Stanford, CA 94305
and the
University of Hawai'i, Mānoa
California State University -- East Bay
plem@carnegiefoundation.org
650-566-5100

Copyright © 2021 - ***The Journal of Writing Assessment*** - All Rights Reserved.