UC Davis
Journal of Writing Assessment

Title
Reliability of National Writing Project's Analytic Writing Continuum Assessment System

Permalink
https://escholarship.org/uc/item/03g148gh

Journal
Journal of Writing Assessment, 6(1)

ISSN
1543-043X

Author
Bang, Hee Jin

Publication Date
2013

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

Reliability of National Writing Project's Analytic Writing Continuum Assessment System

by Hee Jin Bang, National Writing Project

This article presents an investigation of the reliability of a rubric-based writing assessment system, the National Writing Project's (NWP) Analytic Writing Continuum (AWC), which applies both holistic and analytic scoring. Data from double-scored student writing samples collected over several national scoring events (2005 to 2011) were used. First examined was the extent to which scorers trained to apply the AWC tended to agree with each other on the quality of various attributes of student writing (inter-rater agreement rates). Next considerations were how consistently groups of scorers applied the standards of AWC over multiple scoring events (cross-time reliability), and how consistently the attributes of the AWC collectively represented the construct of writing (internal consistency reliability). Finally, generalizability analyses were conducted to determine the degree to which the observed score variances were attributable to two sources of measurement error -- scorers and scoring environment (grade group). Reliability examined from consensus, consistency, and measurement approaches indicate that the AWC assessment system generates highly reliable scoring of both holistic and analytic components of writing. The AWC assessment system includes expert scorers, training procedures, and materials as essential components and serves purposes beyond assessment of writing. It provides a common framework for structuring professional development and coordinating research and evaluation programs, encouraging the growth of professional learning communities and improved understanding of the links between professional development, classroom practice, and student writing performance.

Keywords: writing assessment, scoring rubric, reliability, teacher professional development

---

Introduction

A main objective of this study is to examine the reliability of the National Writing Project's (NWP) Analytic Writing Continuum (AWC) Assessment System (2005, 2010). The NWP serves teachers at all levels (K-16) by providing professional development (PD) and resources for teaching and learning, improving writing instruction, and conducting research on teaching and learning of writing in classrooms. Nearly 200 university-based local writing project sites located in all 50 states each provide high-quality, customized PD for teachers in writing. Since 2003, a cohort of local writing project sites have studied the effects of their programs on instruction and student writing performance. Such sites needed a rigorous scoring system to measure directly the growth of writing performance in an objective, unbiased manner. This need led to the development of the AWC Assessment System by a national panel of experts on student writing and senior NWP researchers, and to the establishment of a national scoring system that would provide impartial assessments of student writing. Originally based on the Six + 1 Trait Writing Model (Culham, 2003), the AWC was developed as an assessment system appropriate for the purposes of evaluating the NWP professional development program and providing data to teachers to inform writing instruction.

Analytic Writing Continuum Assessment System

The AWC provides a general holistic score that indicates the overall quality of writing. The various dimensions of this global construct are captured as the following attributes: (a) Content (central theme or topic, quality and clarity of ideas and meaning); (b) Structure (logical arrangement, coherence, and unity); (c) Stance (perspective communicated through level of formality, style, and tone appropriate for the audience and purpose); (d) Sentence Fluency (rhetorical features, rhythm, and flow crafted to serve the purpose of writing); (e) Diction (precision and appropriateness of the words and expressions for the writing task); and (f) Conventions (usage, punctuation, spelling, capitalization, paragraphing). The scale for the holistic score and each attribute ranges from 1 to 6, and the language used in evaluative statements of the assessment system focuses on the quality of attributes displayed in writing (Swain & LeMahieu, 2012). Such features of the AWC were adopted in favor of the 4-point scale (or a single summary score) and of scoring guidelines that focused on what the writer or writing lacks, which characterizes some large-scale and state assessments. The AWC is appropriate for use across genres of writing and grade levels. This well-defined 6-point scale makes the AWC system sensitive to differences among pieces of writing; the use of six response categories also optimizes the information that can be gleaned from the scoring. (See Preston and Colman (2000) for a discussion on the number of response categories in rating scales in relation to reliability, validity, and discriminating power. See Swain & LeMahieu (2012) for a detailed description of AWC creation.)

The AWC is a well-tested system that has undergone continuous refinements between 2003 and 2011; as of 2012, it has been used at 9 national events to score more than 40,000 student writing samples. Each year, the original AWC design team of writing experts (Range Finding team) reviews the sets of anchor, calibration, and training papers that illustrate the standards for each of the grade group levels: lower elementary (grades 3 & 4), upper elementary (grades 5 & 6), middle school (grades 7 & 8), and high school (grades 9-12). A pool of expert teacher leaders has emerged among those have proven to be reliable scorers and who have developed a deep understanding of the system. These experts now serve as room leaders and table leaders at each scoring event. During training, scorers become familiar with the AWC rubric, examine anchor and practice sets of writing samples, discuss commentaries on the writing samples, and calibrate to a criterion level of performance. During the official scoring, the table leaders

monitor and provide on-going individual support and calibration as needed for scorers at their respective tables.

Literature Review

Theoretical Frameworks

Reliability and validity of holistic and analytic scoring.

Holistic scoring, the primary means of directly assessing writing (Huot, 1990; White, 1994), involves assigning a single score that indicates the overall quality of a text. Raters give one summary score based on their general impression of a text without trying to reduce their judgment to specific set of skills. Analytic scoring involves examining multiple aspects of writing (e.g., content, structure, mechanics, etc.) and assigning a score for each. Thus, holistic scoring focuses primarily on the product, while analytic scoring divides attention to individual components of a product (Finson & Ormsbff, 1998; Rezaei & Lovoron, 2010). Holistic scoring has the advantage of being economical, flexible, and less time consuming (Spandel & Stiggens, 1980), but since each rater can focus on different aspects of writing, the consistency in scoring among scorers is difficult to achieve (Diedrich, French, & Carlton, 1961). Moreover, if raters are basing their judgments on different aspects of writing, some of which may be more pertinent to writing skill than others (e.g., ideas or structure of the text versus the length or handwriting), the validity of the scores is questionable. In contrast, analytic scoring generates several scores that provide information that is potentially useful for guiding instruction and programmatic decisions. Furthermore, examining separate aspects of writing allows for a more comprehensive coverage of the construct, thus increasing the validity of analytic scoring (Quinlan, Higgins, & Wolff, 2009), but it usually requires more time, incurring greater costs (Cooper & Odell, 1977). Different views about which aspects of writing are important to assess and whether certain components should be given more or less weight suggest that consistency in analytic scoring of writing will be challenging to achieve without careful, uniform, training of scorers.

This comparison of holistic and analytic scoring highlights issues of reliability and validity in assessing writing. In broad terms, reliability is the consistency with which an instrument/method produces measurements, while validity is traditionally thought of as the extent to which an instrument/method actually measures what it is meant to measure (i.e., it is an indicator of accuracy). Reliability is a prerequisite for validity, but validity may be observed without reliability; in other words, an instrument may produce consistent scores without measuring what it is meant to measure, but it cannot produce accurate scores unless it also produces consistent scores (Cherry & Meyer, 1993; Moskal & Leydens, 2000). Several different types of reliability and validity exist; evidence of construct validity is critical to making claims about an instrument containing the necessary procedures to truly measure what it is meant to measure (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999; Lyman, 1978; Popham, 1981). Construct validity, however, can only be established over numerous testing situations using the same measure; therefore, this study focuses on the prerequisite - evidence of several types of reliability of the AWC assessment system.

Qualities of writing.

Early research efforts in direct writing assessment have focused on developing reliable holistic scoring procedures, in which reliability is generally referred to as inter-rater agreement (Huot, 1990). Without an explicit guideline, each rater can focus on different qualities of writing and have different judgments about characterizes good writing. Studies on what criteria scorers use to judge writing quality have demonstrated that mechanical correctness, particularly grammar and spelling (Stewart & Grobe, 1979), and superficial features such as length or handwriting (Charney, 1984; Diederich et al., 1961), are highly influential factors that determine how writing is rated (Read, Francis, & Robson, 2005; Ross-Fisher, 2005). Additional research on what influences rater decisions about student writing has demonstrated relationships between holistic scores and specific text characteristics. For example, Nold and Freedman (1977) defined measurable semantic and syntactic features within student essays (e.g., number and development of ideas, complexity, variation, and appropriateness of syntax) to determine which features predicted readers' ratings. They reported that the number of free final modification in a sentence was "indicative of good writing" across genres (p. 173); they also reported high correlations between essay length and quality as well as positive effects of vocabulary on quality. Current research on writing assessment continue to indicate that sophisticated language use (e.g., syntactic complexity, lexical diversity, word frequency) is a strong predictor of ratings indicating writing quality (McNamara, Crossley, & McCarthy, 2010).

Another study conducted by Freedman (1979) investigated the effects of manipulating the categories of content, organization, sentence structure, and mechanics in essays scored on a 4-point holistic scale by twelve teachers. The results showed that while content and organization had the greatest effects, sentence structures and mechanics were also significantly associated with the overall scores. Further, Breland and Jones' (1984) study of essays written for the College Board's English Composition Test and rated by 20 college English professors indicated that overall organization and length of the essay made the greatest contribution to predicting holistic scores. The fact that raters typically assess certain identifiable aspects of writing in order to determine a holistic score suggests that each aspect of writing can be studied individually in greater detail using an analytic scoring system, and that a holistic score alone cannot reveal all the information scorers examined to decide on the final score.

Existing research on writing assessment suggest that holistic and analytic scoring are appropriate for different purposes. The reliability and validity of a scoring procedure needs to be considered in conjunction with issues of practicality, impact of decisions based on the scores, and authenticity of the writing task (Bacha, 2001; Bachman & Palmer, 1996; Weigle, 2002). Holistic scoring may be favored in large-scale assessments or low-stakes placement tests given its cost efficiency. Moreover, since writing involves a broad array of skills and cognitive activities, some would argue that an accurate, valid assessment of writing should only be evaluated as a whole (White, 1985). From an instructional point of view, however, analytic scoring is likely to be favored since the information provided by analytic scoring can help identify areas that would benefit from further instruction as well as areas of strength. Analytic scoring may also be viewed as more valid in the classroom since writing instruction occurs with a focus on one component or skill at a time (Charney, 1984; Flower & Hayes, 1981).

From NWP's standpoint, both holistic and analytic scores should be obtained for a complete assessment of writing. Some assessment systems produce holistic scores that are the arithmetic sum of scores on elemental components of writing, suggesting that the aggregation of individual components amounts to a construct that is identical to what is assessed by the holistic score. The AWC, however, maintains the holistic score and each analytic score as separate, with the premises that overall writing performance can be more or less than the individual disaggregated components, and that the individual components are distinct dimensions of the writing construct (Swain & LeMahieu, 2012). Moreover, some students, especially when writing in a non-native language, can display markedly different levels of competence in various aspects of writing. For example, their understanding and knowledge of content may be much more sophisticated than their sentence structure or mechanics would suggest (Bacha, 2001). Nonetheless, collecting both types of scores raises questions about the independence of those scores; they are bound to be interrelated (Hunter, Jones, & Randhawa, 1996; Singer & LeMahieu, 2012).

Several researchers have indicated that reliability of writing assessment is improved when a rubric is used (Jonsson & Svingby, 2007; Rezaei & Lovorn, 2010); it is possible to achieve high reliability rates with careful training and monitoring throughout the scoring process (Moskal, 2000). Rubrics make explicit the expectations or set of criteria that is used to assess a product; the AWC includes rubrics for holistic and six analytic attributes of writing. In the present study, the author investigated the reliability of the NWP's AWC Assessment System in several ways (Stemler, 2004). The first approach was consensus, i.e., determining the extent to which scorers trained to apply the AWC tend to agree with each other on the quality of various attributes of student writing. A second approach concerned consistency, i.e., how consistently groups of scorers apply the standards of AWC over multiple scoring events, and how consistently the attributes of the AWC collectively represent the construct of writing. The third approach involved measurement, i.e., assessing the extent to which two sources of measurement error - scorers and grade groups - influence the observed scores generated through the application of the AWC system.

Psychometric theory.

In this study, the approach to assessing the reliability of the AWC was guided by principles of classical test theory and generalizability theory (G-theory). Each student/paper has a true score that would be obtained if there were no errors in measurement. In practice, however, one never observes the paper's true score, as the observed score is the combination of the true score plus some error, due to various sources such as occasion, item/forms, and observers/raters (Trochim, 2006). Each observation or measurement is an estimate in an infinitely varied universe of possible measurements (Mushquash & O'Connor, 2006). In G-theory, the potential sources of measurement error are referred to as facets; each facet can have several conditions or levels (e.g., multiple occasions, different items). The various combinations of the facets and levels create "noise," or error variance in measurement of the object being examined (in this case, papers); the averages of the scores from these facets and levels yield the best possible estimates of a paper's true score. G-theory analyses are comprised of identification, estimation, and decomposition of these measurement errors, called variance components; the G-coefficient (ranging from 0 to 1) for each dataset is the true score variance divided by the observed score variance (Cronbach, Linn, Brennan, & Haertel, 1995; Mushquash & O'Connor, 2006). When the G-coefficients are high (at least .80), it is possible to generalize the observed scores across the facets considered in the analyses. The present study focuses on two facets - raters and grade group; these are fixed facets operationalized as a part of the AWC system and thus of particular interest, with implications for NWP research activities.

Research Questions

The specific research questions to be addressed in this study are:

1. To what extent do scorers trained to apply NWP's Analytic Writing Continuum Assessment System (AWC) agree with each other on the quality of various attributes of student writing?
2. How consistently do different groups of scorers apply the standards of AWC over multiple scoring events?
3. How well do the attributes of the AWC capture the construct of writing?
4. To what extent do two sources of measurement error - scorers and grade groups - influence the observed scores produced by application of the AWC?

Methods

Writing Samples

The papers scored were student on-demand writing samples, sent to the NWP prior to scoring events. These writing samples were elicited using open-ended prompts (i.e., not text-based and not genre-specific) from students instructed by teachers across the country who have participated in NWP's professional development programs. Several examples of the prompts are displayed in Appendix A. The NWP staff prepared the papers for scoring, which involved removing all identifying information and creating new IDs and labels to be attached to the papers. From the entire sample of papers to be scored, about 15% were randomly selected and labeled to be scored by two raters, with minimal delay between the readings. (This sample of papers will be referred to as "random set" henceforth.) Raters were considered to be in agreement when they assigned scores that were identical or one point apart. Adjacent scores were viewed as agreement because in reality, writing scores fall along an infinite continuum of scores (Swain & LeMahieu, 2012); however, scorers were forced to choose an integer from 1 to 6. If two scorers determined that a paper was between a 4 and a 5, one scorer may have assigned the lower number, while the other assigned the higher number. In cases where two scores were more than a single point apart, the paper was identified to be adjudicated; the room leader read the writing sample and determined the final score.

In addition to the random set of papers, 498 papers known as the "equating set" were also double-scored. These papers were selected from the first three national scoring conferences (2005-2007). To have a final sample of papers representing all 6 score points across the grade group levels, double scored papers that had 6 or 7 individual sub-scores (six attributes + holistic scores) in agreement or one score point apart from each other were identified. From this pool of papers, approximately 20-25 papers were selected at each score point from 1 through 6 at each grade group level. They were double scored at each subsequent event (2008-2011). At each event, these equating papers and the pre-selected random sample of papers scored by two raters were used to calculate inter-rater agreement rates in real time. These data provide immediate feedback to scorers and inform the re-calibration sessions which occur after each major breaks (overnight, meals). The equating papers also provide data for examining the cross-time reliability of the AWC system, i.e. the consistency with which standards of the AWC are applied across the years.

Results

Inter-rater Reliability

In order to address the first research question - *To what extent do scorers trained to apply NWP's AWC agree with each other on the quality of various attributes of student writing?* - the author examined the data on the random set of 9,000+ double-scored papers that have been read over 9 national scoring events. Agreement rates were used to determine inter-rater reliability, where agreement was defined as identical or adjacent scores, since two raters may have agreed that a paper's holistic score was between 4 and 5, but given the need to choose one integer, one rater assigned a score of 4, while the other rater assigned a score of 5. Nonetheless, the two raters would not have changed their original assessment of the target attribute. Inter-rater agreement rates, calculated using the raw scores assigned by two scorers, have ranged from 89 to 93% across all attributes; an overall agreement rate of 90% has been achieved (Table 1).

Table 1
Inter-rater reliability rates by attribute for 9,129 double-scored papers (22% of total N=40,583) from 2005 to 2011

	Holistic	Content	Structure	Stance	Sentence Fluency	Diction	Conventions
% agreement	93	91	91	89	89	91	89

In addition, to obtain a more detailed view of scoring reliability, the author used the equating set of papers (*N*=498) to examine the number (and percentage) of times when two readers provided identical scores as well as the number of times when the two readers assigned scores that were 1 or more points apart from each other (Table 2). Also examined were the correlations between the scores given by two readers for each attribute across 4 years (Table 3). The high inter-rater agreement rates provide assurances regarding the consistency with which the scorers were applying the standards of the AWC. These scorers, however, represented only one of several sources of variation or error in measurement of writing performance.

Table 2

Frequency and percentage of difference scores between two readers' ratings (total N = 498)

	Holistic	Content	Structure	Stance	Sentence Fluency	Diction	Conventions
Difference in scores				2008			
0	216	210	212	169	204	216	208
	43.4%	42.2	42.8	33.9	41.0	43.4	41.8
1	239	237	214	239	223	215	223
	48.0%	47.6	43.0	48.0	44.8	43.2	44.8
2	42	46	60	82	67	63	61
	8.4%	9.2	12.1	16.5	13.5	12.7	12.3
3	1	5	11	8	4	4	6
	0.2%	1.0	2.2	1.6	0.8	0.8	1.2
				2009			
0	244	247	250	222	250	240	245
	49.0%	49.6	50.2	44.6	50.2	48.2	49.2
1	222	212	216	207	200	215	212
	44.5%	42.3	43.4	41.6	40.2	43.2	42.6
2	28	35	26	63	43	40	36
	5.6%	7.0	5.2	12.7	8.6	8.0	7.2
3	4	4	6	6	5	3	5
	0.8%	0.8	1.2	1.2	1.0	0.6	1.0
				2010			
0	220	209	224	190	199	204	199
	44.2%	42.0	45.0	38.2	40.0	41.0	40.0
1	230	240	217	234	241	250	222
	46.2%	48.2	43.6	47.0	48.4	50.2	44.6
2	43	44	52	63	52	38	67
	8.6%	8.9	10.4	12.7	10.4	7.6	13.5
3	5	5	5	11	5	5	7
	1.0%	1.0	1.0	2.2	1.0	1.0	1.4
4	0	0	0	0	1	1	3
	0%	0	0	0	0.2	0.2	0.6
				2011			
0	274	272	272	230	258	265	265
	55.0%	54.6	54.6	46.2	51.8	53.2	53.2
1	221	224	223	266	237	230	232
	44.4	45.0	44.8	53.4	47.6	46.2	46.6
2	3	2	3	2	3	3	1
	0.6%	0.4	0.6	0.4	0.6	0.6	0.2

Table 3

Pearson r coefficient between two scorers' ratings by attribute for 498 papers in the equating set

	Holistic	Content	Structure	Stance	Sentence Fluency	Diction	Conventions
2008	.72***	.69***	.63***	.63***	.66***	.64***	.66***
2009	.71***	.71***	.71***	.66***	.70***	.68***	.71***
2010	.66***	.65***	.65***	.62***	.63***	.65***	.62***
2011	.80***	.80***	.79***	.80***	.81***	.79***	.83***

***$p < .001$

Cross-time Reliability

The second research question (*How consistently do different groups of scorers apply the standards of AWC over multiple scoring events?*) focuses on another source of variation - time, or year of scoring. Analyses to assess cross-time reliability of the AWC are based on the dataset of 498 equating papers (described above) that have been scored by two independent raters in 2008, 2009, 2010, and 2011. All four grade groups are represented in the sample: lower elementary (3^{rd}-4^{th} grades, n=159), upper elementary (5^{th}-6^{th} grades, n=82), middle school (7^{th}-8^{th} grades, n=143), and high school (9^{th}-12^{th} grades, n=114). The means and standard deviations of all scores are presented in Table 4.

Table 4

Means and Standard Deviations: Sum of Two Scores (range 2-12) on Each Attribute for Papers (N=498) Scored by Two Raters Across Four Scoring Conferences

	2008	2009	2010	2011
	Mean (SD)			
Holistic	6.32 (2.27)	6.05 (2.10)	6.18 (2.10)	6.10 (2.06)
Content	6.39 (2.27)	6.15 (2.12)	6.20 (2.09)	6.13 (2.05)
Structure	5.86 (2.22)	5.70 (2.09)	5.84 (2.10)	5.77(1.99)
Stance	6.65 (2.37)	6.30 (2.25)	6.45 (2.24)	6.38 (2.23)
Sentence Fluency	6.33 (2.27)	6.10 (2.18)	6.20 (2.13)	6.26 (2.16)
Diction	6.34 (2.14)	6.13 (2.05)	6.17 (2.09)	6.16 (2.04)
Conventions	6.11 (2.23)	6.04 (2.07)	6.10 (2.26)	6.04 (2.23)
KMO measure of sampling adequacy (values close to 1 indicate adequate representation in sample)	.932	.923	.931	.942

To assess how consistently the AWC standards were applied over the years, the author examined the correlation coefficients among all the attribute scores across years. As shown in Table 5, the correlations range from .73 to .83. In addition, the frequency distributions of the scores from each of the four grade groups over all four years were graphed on a single coordinate plane, one for each of the 6 attributes and 1 holistic score. To increase the sensitivity of the scale, the two scores assigned to each attribute were summed, creating a range of scores 2 (1+1) through 12 (6+6). The frequencies of these scores were determined for all 498 papers and examined by grade group level. These frequency distribution graphs (ogives) show the extent to which scores vary over time (year of scoring) and across different grade group levels (i.e., scoring rooms). Graphs presented in Figure 1 show the cumulative percent of Holistic scores for each grade group across four years of scoring; ogives for all other attributes in each grade group resemble that of the Holistic score. These graphs indicate that there is small variance in observed scores across the years; even in the central part of the distribution where the most variation is expected (scores 6, 8), the observed scores are quite consistent. This consistency is confirmed by correlational analyses performed using the summed scores (sum of two scores given by two raters), which demonstrate for each attribute the relationships of scores across multiple years.

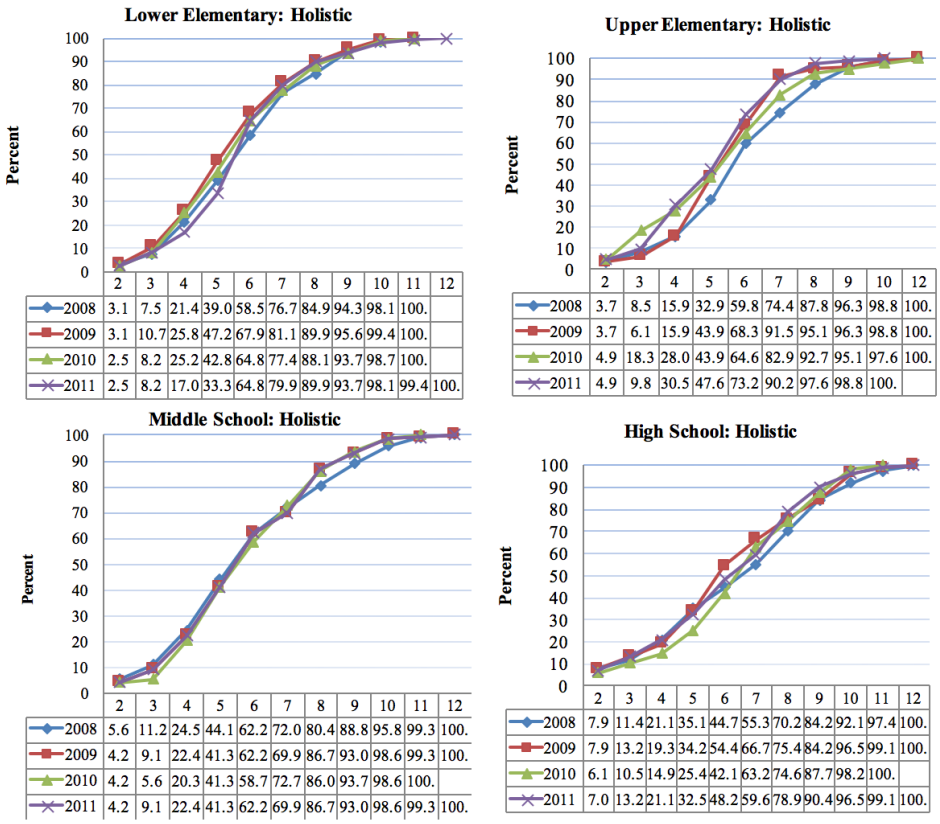
Table 5

Correlations (Pearson r) of scores across years 2008-2011 by attribute

Holistic	2008	2009	2010	2011
2008	1	.81**	.80**	.81**
2009		1	.84**	.77**
2010			1	.78**
2011				1
Content				
2008	1	.79**	.78**	.79**
2009		1	.81**	.75**
2010			1	.78**
2011				1
Structure				
2008	1	.79**	.78**	.79**
2009		1	.82**	.76**
2010			1	.78**
2011				1
Stance				
2008	1	.74**	.76**	.75**
2009		1	.80**	.73**
2010			1	.75**
2011				1
Sentence Fluency				
2008	1	.78**	.76**	.80**
2009		1	.82**	.75**
2010			1	.78**
2011				1
Diction				
2008	1	.77**	.75**	.77**
2009		1	.82**	.75**
2010			1	.76**
2011				1
Convention				
2008	1	.79**	.75**	.77**
2009		1	.77**	.76**
2010			1	.77**
2011				1

**$p < .01$

Figure 1. Cumulative percent distributions of Holistic scores (2008-2011) by grade group level.



Internal Consistency Reliability

The third research question (*How well do the attributes of the AWC capture the construct of writing?*) involved assessing the internal consistency reliability of the AWC scores from each year by examining the intercorrelations among the attributes and estimating Cronbach's alpha, an index of how well the attributes measure the underlying construct of writing ability. Prior to calculating the reliability indices, factor analyses were conducted to investigate the degree to which individual attributes capture various dimensions of the writing construct. For these analyses, the summed scores for each attribute assigned by two raters to the papers in the equating set were used. Holistic scores were excluded to minimize the multicollinearity that arises when multiple attributes are measures of the same construct. (Correlations among the 6 attributes and Holistic scores within each year ranged from .82 to .97 across 4 years.) Factor analyses were specified in SPSS to generate maximum likely estimates, using oblimin rotation on the 6 attribute scores. The results consistently produced a clear 1-factor solution (see Figure 2), with no other factors obtaining an eigenvalue greater than 1. Approximately 88% of the total variance in scores was explained by the single factor, and factor loadings of attributes ranged from .87 to .96 across 4 years, indicating multicollinearity (Table 6). The proportion of each attribute score explained by the single factor ranged from .75 to .93 (Table 7).

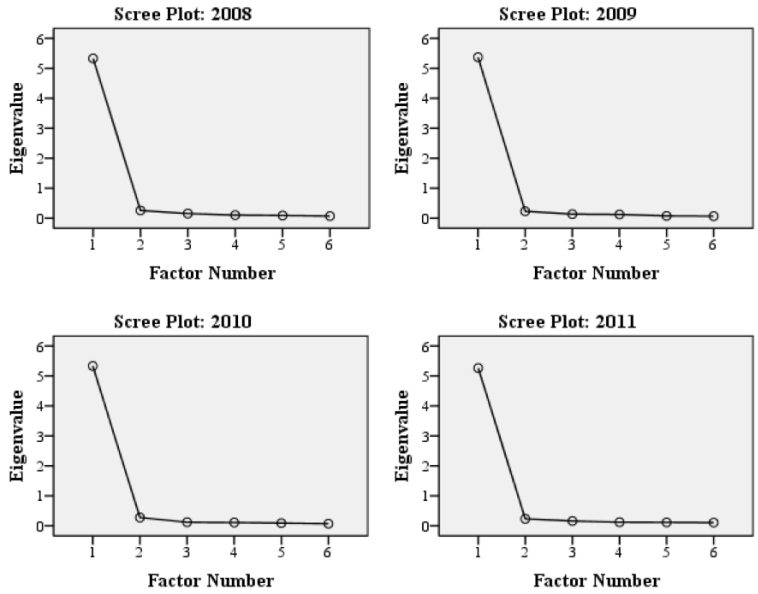
Table 6
Factor Loadings of Attributes

	2008	2009	2010	2011
Content	.962	.954	.958	.942
Structure	.953	.936	.954	.927
Stance	.934	.915	.922	.929
Sentence Fluency	.941	.935	.944	.939
Diction	.916	.942	.939	.928
Conventions	.872	.929	.867	.874
Eigenvalues	5.329	5.371	5.333	5.263
Percent of total variance	88.822	89.510	88.887	87.719

Table 7
Communalities: Extraction (proportion of each attribute score explained by the underlying factor)

	2008	2009	2010	2011
Content	.926	.910	.918	.888
Structure	.908	.876	.911	.860
Stance	.872	.838	.849	.864
Sentence Fluency	.886	.874	.891	.881
Diction	.840	.887	.889	.861
Conventions	.760	.862	.752	.764

Figure 2. Scree plots generated from factor analyses of Content, Structure, Stance, Sentence Fluency, Diction, and Conventions scores across years.



All attribute scores from each year were then submitted for reliability analyses. Results showed that the internal consistency of the AWC scores is remarkably high, with item-total correlations ranging from .86 to .94. The average alpha coefficient for the scale across the 4 years is .974 (Table 8). None of the attributes across 4 years were identified as items to be excluded in order to improve the alpha coefficient, with the exception of 2010 Conventions, but its removal had only miniscule effect (.002) on the Cronbach's alpha for that year. Although the high communalities observed in the factor analyses suggest some degree of redundancy, NWP refrains from eliminating any of the attributes because the AWC Assessment System serves functions beyond simple measuring of writing. It informs teachers' classroom practice by identifying strengths and weaknesses of writing performance to highlight specific areas of writing that deserve targeted instruction.

Table 8
Reliability and Scale Statistics

	2008	2009	2010	2011
Cronbach's α (N=6 items)	.975	.976	.974	.971
Scale Mean (SD)	37.68 (12.72)	36.41 (12.06)	39.96 (12.15)	36.74 (11.89)

Generalizability Theory Analyses

The fourth research question (*To what extent do two sources of measurement error - scorers and grade groups - influence the observed scores?*) was addressed through G-theory analyses, which allow for identification of sources of variance in a given design that have the greatest influence on the measured observations. It is possible to estimate how much the measure for an object of study differs from the average score that would be obtained by measuring the object under the entire universe of possible conditions. Two major sources of measurement error were selected to be investigated (scorers and grade groups). *EduG*, a software program designed to perform generalizability analyses, was used. The program is based on analysis of variance designs that are complete and balanced (Educan, Inc., 2010); the estimated variance components in the ANOVA results are used to compute the generalizability parameters. The dataset of 498 papers comprised of different numbers of papers at each grade group; therefore, stratified random sampling was used to select 82 papers (the smallest total number of papers for a grade group, i.e., upper elementary) from each of the other three grade groups. This process resulted in a dataset of 328 papers for use in the G-theory analyses.

The G-study involved specifying three designs. The observation design described the structure of the dataset in terms of facets and levels involved; the estimation design indicated the size of the population for each facet; and, the measurement design distinguished the various facets that contribute to the true score variance, the instrumentation facets, and potential error variance (Educan, Inc., 2010). In *EduG*, estimates of the variance of the facets under study are compared with an estimate of the total variance, taking into account the sampling involved in each of the facets (fixed or random). The generalizability estimates (coefficient G) that are calculated indicate the proportion of the true score variance in the total expected observed score variance, (i.e. the percentage of the observed score variance that is in fact the true score variance).

The objective of the G-theory analyses was to identify the degree to which scorers (or raters) and grade groups influence the observed scores. The observation design was defined using single-letter facet identifiers (R for raters or scorers, G for grade group, and P for papers) and the number of levels that have been observed for each facet (2 levels in raters, 4 groups in grade groups, and 328 in papers). In specifying the estimation design, which affects the degree to which results can be generalized, the facet universe for Rater was fixed at 500 (approximate number of teacher leaders available to serve as a rater); the universe of Grade group was

fixed at 4 (lower elementary, upper elementary, middle school, and high school); and an infinite population was indicated for Paper. The measurement design defined Paper and Grade group as differentiation facets (the facets being differentiated) and Rater was identified as an instrumentation facet. The identifier for a nested facet includes the identifier of nested facet, separated by a colon (e.g., P:G indicates that Paper facet is nested in Grade group facet).

Table 9 displays a summary of the ANOVA and G-study analyses for each of the attribute scores. The columns under ANOVA show sources of variance, the proportion of the variance of individual scores that is attributable to each variance source, along with the standard errors associated with each of the estimated variance components. These ANOVA results are used in the G-study, which entails assessing the quality of measurement for the differentiation facets chosen for the present analyses. In the first column under G-study, the error variance, along with the impact of each source of variance on the variance of errors is shown. Standard errors allow one to determine the confidence interval around the true mean score of the object of measurement; coefficient-G's indicate the quality of the overall design. The ANOVA results in Table 9 indicate that across all attribute scores, approximately 80% of the variance of the observed scores is attributable to the Paper (nested in Grade group), while the remaining 20% of the variance is attributable to the interaction of Rater and Paper (nested in Grade group). The G-study results show that about 98% of the error variance is attributed to the Rater by Paper interaction (within the Grade group). Coefficient-G's range from.88 to .91; these indicate a very satisfactory level of precision in the overall design.

Table 9
Summary of Analysis of Variance and G-study for Each Attribute Score

	ANOVA		G-study (Measurement design PG/R)		
Attribute score	Source of variance	% of variance of scores attributable to variance source (SE)	error variance (% of error attributable to variance source)	Standard Error	Coefficient-G[1]
Holistic	R	0.0 (.00)	.00 (0.2)	.35	.89
	G	---	---		
	P:G	80.1 (.04)	1.01		
	RG	0.3 (.00)	.00 (1.7)		
	RP:G	19.5 (.01)	.12 (98.3)		
Content	R	0.1 (.00)	.00 (0.3)	.35	.89
	G	---	---		
	P:G	80.8 (.04)	1.02		
	RG	0.3 (.00)	.00 (1.7)		
	RP:G	18.8 (.01)	.12 (98.3)		
Structure	R	0.2 (.00)	.00 (0.9)	.35	.89
	G	---	---		
	P:G	79.3 (.04)	.95		
	RG	0.3 (.00)	.00 (1.4)		
	RP:G	20.2 (.01)	.12 (98.6)		
Stance	R	0.2 (.00)	.00 (0.9)	.36	.88
	G	---	---		
	P:G	79.0 (.05)	1.06		
	RG	0.4 (.00)	.00 (1.9)		
	RP:G	20.4 (.01)	.14 (97.3)		
Sentence Fluency	R	0.1 (.00)	.00 (0.8)	.34	.90
	G	---	---		
	P:G	81.2 (.05)	1.04		
	RG	0.3 (.00)	.00 (1.5)		
	RP:G	18.4 (.01)	.12 (98.5)		
Diction	R	0.0 (.00)	---	.32	.91
	G	---	---		
	P:G	83.3 (.04)	1.04		
	RG	0.2 (.00)	.00 (1.1)		
	RP:G	16.5 (.01)	.10 (98.9)		
Conventions	R	0.2 (.00)	.00 (0.8)	.36	.89
	G	---	---		
	P:G	79.8 (.04)	1.01		
	RG	0.2 (.00)	.00 (1.0)		
	RP:G	19.8 (.01)	.12 (99.0)		

[1] Relative coefficient-G's are reported (as opposed to absolute coefficient-G's), which take into account the sources of variance affecting the relative scale of measurement.

Discussion

This study provides evidence that the AWC assessment system yields highly reliable measurements of both holistic and analytic components of writing. The AWC assessment system has shown a degree of reliability and validity higher than many standardized

writing assessments available. Widely used multiple-choice tests that are machine-scored may have very high reliabilities, but these indirect measures of writing do not actually measure writing ability and tend to encourage teachers to place emphasis on narrow aspect of writing (e.g., grammar, usage); they are thus limited in validity (Murphy, 2008; White, 1985). Tests such as the SAT and GRE have actual writing components, but they are single, timed essays that are scored holistically only; general impressions tend to depend on the rater's definition of quality.

Using both holistic and analytic procedures, the AWC is a robust instrument that measures all the important aspects of writing, consistently across multiple raters and scoring sessions. The AWC Assessment System encompasses all aspects of scoring, including selection of scorers, anchor papers, training papers, procedures for conducting training and calibration, actual scoring of papers, and adjudications. The global design of this system, involving two raters and four grade groups, has been demonstrated to produce scores that are of very satisfactory precision. In psychometric terms, the data clearly support a one-factor solution. The high communalities observed in factor analyses and alpha coefficients obtained from internal consistency reliability analyses suggest some degree of redundancy, and it can be tempting to use the holistic scores only or a smaller set of attributes in efforts to minimize time and cost of assessing writing. If the sole purposes of an assessment were to sort, distribute, and rank performance, the collapsing of scores (or even the use of the holistic scale alone), would be perfectly reasonable. However, the purposes of the AWC Assessment System extend beyond mere sorting and ranking of writing performance; to maximize the capabilities of the AWC Assessment System, it should be used as a multi-dimensional scale.

It is widely recognized that assessment has a strong influence on instruction (Cheng, Watanabe, & Curtis, 2003). The AWC's explicit descriptions of various writing attributes and the criteria by which each attribute will be judged make the assessment transparent to both teachers and students (Johsson & Svingby, 2007). The AWC can be used by students to give feedback to peers' writing, and research has demonstrated positive effects of exchanging feedback on students' learning (Dochy, Segers, & Sluijsmans, 1999). Students can also self-assess their writing and identify areas to further develop, engaging in metacognition as they work on honing specific skills (Falchikov & Goldfinch, 2000; Hafner & Hafner, 2003). Such use of the AWC by students will enable them to identify and internalize the characteristics of quality writing; the transparency of assessment can also motivate them to learn the craft of writing (Cauley & McMillan, 2010; Goslin, 2003).

For teachers, the AWC rubric provides a clear set of instructional goals. It focuses the content of classroom instruction and offers a structure for systematically addressing specific aspects of writing. The rubric applied to classroom work can point to students' strengths and weaknesses of writing performance, as well as growth in particular areas. The fact that there are significant differences between the means on the various scales consistently and across years suggests that even if the auto correlations are high and each scale orders performances similarly, they arrive at different location parameters (i.e., levels of performance). Thus, the AWC can provide diagnostic information that shape teachers' instructional approach, highlight specific areas of writing that deserve targeted instruction, and lead teachers to focus instruction to address the needs of individual students or class groupings. By promoting examinations of various components of writing instruction, the AWC also serves as a tool in professional development on teaching writing and encourages teachers to reflect on their practice and monitor the effectiveness of particular instructional approaches (Schamber & Mahoney, 2006; Shaw, 2004). The AWC therefore provides a common language and metric around which professional development can be structured, encouraging the growth of professional communities, supporting teachers' growth as writers and as teachers of writing, and improved student learning outcomes (Swain & LeMahieu, 2012).

Another compelling reason for using both analytic and holistic scoring comes from a coordinated program of research, conducted by NWP from 2003 to 2011, to assess the impact of professional development work on teachers' classroom practice and student writing performance. NWP worked closely with local Writing Project sites to design a series of 18 experimental and quasi-experimental studies, 17 of which examined inservice professional development programs provided by NWP teacher-leaders. The results of analyses conducted by independent evaluators who played no role in leading the programs indicate that student results are consistently strong and favorable in specific aspects of writing for which the NWP is best known, such as development of ideas (content) and organization (structure). Analytic scoring of writing and measuring change in each attribute enabled researchers to identify program components that produced the observed change in particular aspects of writing. If the student writing samples had been scored using the holistic measure only, changes may have gone undetected, and making links between program components and outcomes would have been difficult.

Future Directions

The results of the factor analyses presented in this paper (i.e., high factor loadings and communalities for all attributes) prompt NWP researchers to consider alternative ways of examining the factors underlying individual attributes. Descriptions of each attribute in the AWC rubrics indicate that substantial differences exist among them (in other words, face validity or content validity is observed); yet the results of the factor analyses appear to blur the distinctiveness of attributes.

In addition, future validation efforts of the AWC will examine other sources of error variance such as prompts, year of scoring, and trainers. In particular, the effects of prompts that elicit specific genre of writing (e.g., persuasive, informative, narrative) will be

investigated. The coefficient-G's will reveal the sources of variation that contribute the most or least to the error variance and inform decisions (through D-study analyses) about ways to optimize the design and improve the cost-efficiency of scoring without compromising the technical rigor of the AWC Assessment System.

Author Note

Hee Jin Bang is a Senior Research Associate at the National Writing Project. Her current research interests include English Language Learners' (ELL) development of writing skills, writing instruction for ELLs, and cross-national contexts of youth development and civic engagement. She has published on homework in the academic lives of immigrant origin youth, and academic, social, and psychological adaptation of youth growing up in transnational families. Her work in the areas of teacher professional development and evaluation of teacher preparation programs is centered on the goal of understanding the teachers' role in preparing an increasing number of ELLs for postsecondary education and the workforce.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System, 29*, 371-383.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Breland, H. M. & Jones, R. J. (1984). Perceptions of writing skills. *Written Communication, 1*, 101-09.

Cauley, K. M. & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 83*, 1-6.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English. 18*(1), 65-81.

Cheng, L., Watanabe, Y. J., & Curtis, A. (2004). *Washback in Language Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cherry, R. D. & Meyer, P. R. (1993). Reliability issues in holistic assessment. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109-141). Cresskill, NJ: Hampton Press.

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

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). Generalizability analysis for educational assessments. UCLA Center for the Study of Evaluation & The National Center for Research on Evaluation, Standards, and Student Testing.

Culham, R. (2003). *6 + 1 Traits of Writing*. New York: Scholastic Professional Books.

Diederich, P. B., French, J. W. & Carlton, S. T. (1961). Factors in the judgment of writing quality. Princeton: Educational Testing Service, ETS RB NO 61-15.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: a review. *Studies in Higher Education, 24*, 331-350.

Educan Inc. (2010). *EduG User Guide*. Edumetrics - Quality of measurement in education. Retrieved from http://www.irdp.ch/edumetrie/documents/EduGUserGuide.pdf.

Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70,* 287-322.

Finson, K. D. & Ormsbff, C. K. (1998). Rubrics and their use in inclusive science. *Intervention in School and Clinic, 34*(2), 79-88.

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

71(3), 328-338.

Goslin, D. A. (2003). *Engaging minds: Motivation and learning in America's schools*. Lanham, MD: The Scarecrow Press, Inc.

Grobe, C. (1981). Syntactic maturity, mechanics and vocabulary as predictors of quality ratings. *Research in the Teaching of English, 15*, 75-88.

Hafner, J. C. & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*, 1509-1528.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201-213. DOI: http://www.jstor.org/stable/358160.

Hunter, D. M., Jones, R. M., & 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.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130-144.

Lyman, H. B. (1978). *Test Scores and What They Mean* (3rd ed). Englewood Cliffs: Prentice. McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Writing Communication, 27*(1), 57-86.

Moskal, B. M. 92000). Scoring rubric: what, when and how? *Practical Assessment, Research & Evaluation, 7*(3). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=3.

Moskal, B. M. & Leydens, J. A., (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=10.

Murphy, S. (2008). Some consequences of writing assessment. In A. Havnes & L. McDowell (Eds.) *Balancing dilemmas in assessment and learning in contemporary education*. New York: Routledge.

Mushquash, C. & O'Connor, B. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542-547.

National Writing Project. (2005, 2010). The Analytic Writing Continuum: A comprehensive writing assessment system. University of California, Berkeley; Berkeley, CA: National Writing Project.

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

Nold, E. W., & Freedman, S. W. (1977) An analysis of readers' responses to essays. *Research in the Teaching of English, 11*, 164-74.

Popham, J. W. (1981). *Modern Educational Measurement*. Englewood: Prentice.

Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1-15.

Quinlan,T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater® scoring engine. Princeton, NJ: Educational Testing Services.

Read, B., Francis, B., & Robson, J. (2005). Gender, bias, assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment and Evaluation in Higher Education, 30*(3), 241-260.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*, 18-39.

Ross-Fisher, R. L. (2005). Developing effective success rubrics. *Kappa Delta Pi, 41*(3), 131-135.

Schamber, J. F. & Mahoney, S. L. (2006). Assessing and improving the quality of group critical thinking exhibited in the final projects of collaborative learning groups. *Journal of General Education, 55*, 103-170.

Shaw, J. (2004). Demystifying the evaluation process for parents: Rubrics for marking student research projects. *Teacher Librarian, 32*, 16-19.

Singer, N.R. & LeMahieu, P. (2012). The effect of scoring order on the independence of holistic and analytic scores. *Journal of Writing Assessment, 4*.

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

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4.

Stewart, M. F., & Grobe, C. H. (1979). Syntactic maturity, mechanics, vocabulary and teachers' quality ratings. *Research in the Teaching of English, 13*, 207-215.

Swain, S. & LeMahieu, P. (2012). Assessment in a culture of inquiry: The story of National Writing Project's Analytic Writing Continuum. In N. Elliot and L. Perelman (Eds). *Writing assessment in the 21st century: Essays in honor of Edward White*. New York: Hampton Press.

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

White, E. M. (1985). Holisticism. *College Composition and Communication, 35*(4), 400-409.

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

Appendix A

Sample of prompts that elicited writing samples in the equating set

- If you could travel back in time or into the future, what time period would you pick and where would you go?

- The library is sponsoring a short story contest. The only rule is that every story must include the following sentence: It was all the fault of the baby elephant.

- We all have friends who are very important to us. Think about one friend who is special to you. Write to inform your teacher about the characteristics of this friend that are most important to you. Tell how your friend exhibits these characteristics. Be sure to include a lot of details about this friend so that your teacher understands the characteristics that make this friend special to you.

- Is there a special event such as a concert, sports event, multicultural night, school play, family event, nature hike, or other event you have recently attended that you especially enjoyed? Write a review of this event for your school newspaper or a review that could be printed in the youth section of your local newspaper. Include a title, a brief summary of the event and its participants and an explanation of what made the event so entertaining or interesting to you.

- Your local TV broadcasting station is having a contest to determine which TV shows to take off the air for the new season. Write a well-organized, multi-paragraph essay to persuade the broadcaster to keep your favorite show on the air. Be sure to include specific and relevant details to support your opinion.

- Many people would like to reduce the legal age for driving to 14 (it is now 16 in most states). What is your opinion? Support your position with a convincing argument that could be printed in the editorial section of the newspaper.

- You borrow an article of clothing from someone for a special occasion. Much to your horror, disaster strikes, and you ruin the borrowed item. Write about what happened and why you handled the situation the way you did.

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