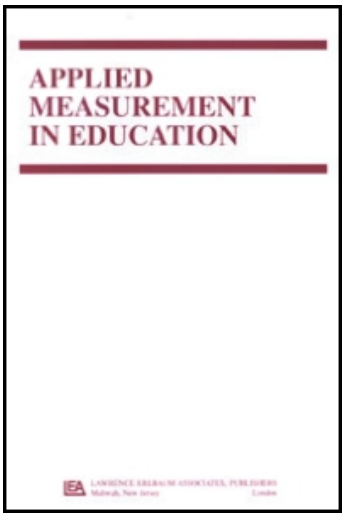

Applied Measurement in Education
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653631

The Critical Role of Anchor Paper Selection in Writing Assessment

Sharon E. Osborn Popp [a]; Joseph M. Ryan [a]; Marilyn S. Thompson [b]
[a] College of Teacher Education and Leadership, Arizona State University, [b] Division of Psychology in Education, Mary Lou Fulton College of Education, Arizona State University,

Online Publication Date: 01 July 2009

To cite this Article Osborn Popp, Sharon E., Ryan, Joseph M. and Thompson, Marilyn S.(2009)'The Critical Role of Anchor Paper Selection in Writing Assessment',Applied Measurement in Education,22:3,255 — 271
To link to this Article: DOI: 10.1080/08957340902984026
URL: http://dx.doi.org/10.1080/08957340902984026

APPLIED MEASUREMENT IN EDUCATION, 22: 255–271, 2009

ISSN: 0895-7347 print / 1532-4818 online
DOI: 10.1080/08957340902984026

The Critical Role of Anchor Paper Selection in Writing Assessment

Sharon E. Osborn Popp and Joseph M. Ryan
College of Teacher Education and Leadership
Arizona State University

Marilyn S. Thompson
Division of Psychology in Education
Mary Lou Fulton College of Education
Arizona State University

Scoring rubrics are routinely used to evaluate the quality of writing samples produced for writing performance assessments, with anchor papers chosen to represent score points defined in the rubric. Although the careful selection of anchor papers is associated with best practices for scoring, little research has been conducted on the role of anchor paper selection in writing assessment. This study examined the consequences of differential selection of anchor papers to represent a common scoring rubric. A set of writing samples was scored under two conditions—one using anchors selected from within grade and one using anchors selected from across three grade levels. Observed ratings were analyzed using three- and four-facet Rasch (one-parameter logistic) models. Ratings differed in magnitude and rank-order, with the difference presumed to be due to the anchor paper conditions and not a difference in overall severity between the rater groups. Results shed light on potential threats to validity within conventional context-dependent scoring practices and raise issues that have not been investigated with respect to the selection of anchor papers, such as the interpretation of results at different grade levels, implications for the assessment of progress over time, and the reliability of anchor paper selection within a scoring context.

Correspondence should be addressed to Sharon E. Osborn Popp, College of Teacher Education and Leadership, P.O. Box 37100, Phoenix, AZ 85069-7100. E-mail: osbornpopp@asu.edu

Direct assessments of writing performance are included in many large-scale testing programs, including 31 state assessment programs as of 1999–2000 (Goertz & Duffy, 2001). Assessments of writing performance often carry high-stakes consequences despite validity concerns (Gordon, Engelhard, Gabrielson, & Bernknopf, 1996; Mehrens, 1992). One threat to validity in writing performance assessment is construct irrelevant variance due to deficiency in the quality and consistency of scoring. Messick (1995) identified the structural integrity of the scoring framework as a critical aspect of validity for performance assessments. An essential feature of the scoring framework is the scoring rubric, which is routinely applied to evaluate the quality of written samples produced for writing performance assessments. The performance levels defined within the rubric are made explicit through anchor papers, which are the sets of papers chosen to represent score points along the rubric's scale and guide scoring. The purpose of this study was to investigate the role of anchor papers in a direct assessment of writing performance. After examining the consequences of differential selection of anchor papers used to represent a common writing rubric, we discuss potential threats to validity within conventional scoring practices and introduce implications for further action and research.

Anchor papers, also known as benchmark papers, exemplars, or range finders, are the writing samples chosen to define levels of performance in the scoring rubric. The chosen anchor papers operationalize the concepts described in the language of the scoring rubric. They define the standards of performance for a given assessment and serve as the rubric's surrogate reference points, against which all samples are judged. Anchor papers are usually selected in a process called range finding. During range finding, the rubric is studied carefully and a set of students' papers are reviewed to identify papers that exemplify the various score points on the rubric.

The consistent application of the scoring rubric is considered essential to the validity and meaningful interpretation of scores for performance assessments (see e.g., Brennan & Johnson, 1995; Messick, 1995). The main elements involved in the scoring of writing performance assessments are rubrics, anchor papers, and raters. Attention has been given to many aspects of scoring in writing performance assessment, such as choice of rubric (Huot, 1990; Novak, Herman, & Gearhart, 1996; Roid, 1994), quality of rater training and experience (Freedman, 1981; Huot, 1993; Wolfe, Kao, & Ranney, 1998), conduct of rating sessions, (White, 1985), use of sequential versus spiral scoring/training models (Moon & Hughes, 2002), order effects in scoring (Hughes & Keeling, 1984), rater scoring accuracy (Engelhard, 1996; Quellmalz, 1984), resolution methods for rater disagreement (Cherry & Meyer, 1993; Johnson, Penny, & Gordon, 2000; Myford & Wolfe, 2002), and systematic patterns of rater errors (Coffman, 1971; Engelhard, 1994; Saal, Downey, & Lahey, 1980; Thorndike, 1920). The particular anchor papers chosen to represent levels of performance in the rubric would also appear

to be highly related to score outcome. However, research regarding the role of anchor paper selection in scoring direct writing assessments is surprisingly limited.

Mullis (1984) observed that procedures used to define score points in writing assessments may vary, ranging from "analysis of the corpus of papers to be scored to establish relative definitions to using absolute definitions established prior to collecting students' responses" (p. 16). Some holistic scoring procedures explicitly call for selecting anchor papers from the particular set of samples to be scored before establishing a scoring scale, so that the rubric is inferred rather than imposed (Daiker & Grogan, 1985; Odell & Cooper, 1980). However, many current large-scale writing assessments employ analytic rubrics with predefined scale points against which the writing is judged; anchor papers are selected to reflect each scale point.

State education departments and school districts document the use of carefully selected anchor papers as a necessary and standard step in the training and scoring procedures for writing assessment (see e.g., Driscoll, 1996; Nevada Department of Education, 2000; Washington Office of the State Superintendent of Public Instruction, 2001). Arizona's Instrument to Measure Standards official scoring materials for writing assessment (Arizona Department of Education, 2005) state that "The same rubric will be used for all grade levels, with developmental differences taken into consideration. The characteristics of effective writing do not change depending on what grade you are in, only the level of sophistication changes" (p. 1). Regarding scale point interpretation, the scoring guide explains that establishing anchor papers through range finding is critical in interpreting points on the scale and maintaining scoring consistency.

The selection of anchor papers to reflect the rubric's score points is widely presumed to be associated with best practices that ensure assessment quality and high inter-rater reliability through clearly illustrating the intent of the rubric to the raters. However, little research has been conducted to explore whether and how the selection of anchor papers to represent the intent of an analytic rubric improves the reliability and validity of assessment. One school district reported the introduction of anchor papers as one part of a strategy to improve their writing assessment, with results indicating no improvements in the degree of evidence to support increased reliability or validity (Fenton, Straugh, Stofflet, & Garrison, 2000). The rescoring of anchor papers for a state writing performance assessment was reported for a study conducted to investigate unusually low performance for a particular year (Zhang, 2000). Zhang reported that consistency in the rescoring of the same anchor papers varied from 73% to 92%, with lower rates of consistent scoring associated with higher grade levels.

We sought to investigate whether, and to what extent, anchor paper selection influences the ratings of students' writing. In this study, we examine the consequences of anchor paper selection under two scoring conditions that differ with respect to grade level. The comparison of results scored with anchor papers

selected from a single grade level and from across multiple grade levels addresses the issue of whether the definitions of scale points in an analytic rubric can be expected to remain absolute when interpreted against sets of writing that represent different ranges of writing quality. If the same rubric is applied to different sets of writing samples to be scored, there is a reasonable expectation that it will be applied consistently, regardless of the range of quality manifested in the samples.

This study considers the role of anchor papers with respect to two main issues. One issue is whether the score points in a rubric can be represented reliably in anchor paper selection to produce consistent scoring. If factors other than the rubric are considered in scoring, such as grade-level expectations or specific writing features associated with different discourse modes, then the role of the anchor papers is to represent those additional expectations in addition to the expectations defined in the rubric. The selection of anchor papers thus depends on the scoring context, which in practice, depends strongly on the sample of papers from which the anchors are chosen. The second issue is how the selection of different anchor papers within different contexts, in this case different grade-level assessments, transforms the standards on which the evaluation framework is based. The writing rubric employed in this study defines a broad domain of writing that can describe the performance of emergent, novice writers through highly proficient, expert writers. The writing responses of many students in lower grade levels would be expected to reflect the lower scale points of the rubric while responses of many students in higher grades would be expected to reflect higher scale points. When applied to the writing responses from a conventional, single grade-level assessment, does the process of range finding consider the broad construct of writing implied in the rubric, or does the process consider the body of papers at hand to define the standards within the context of grade-level expectations? The application of the rubric to anchor paper selection within each grade level has implications for the assessment of progress over time and the interpretation of results at different grade levels.

METHOD

Sample and Data Collection Design

The writing samples were produced for a district writing assessment by students in grades 3, 5, and 8 from a moderately large metropolitan school district. All grade 3 students responded to the following narrative mode prompt: “Think of something you have done, a special place you have been, or a special person you have known that has created a memory for you. Describe your feelings and why it was important to you.” Approximately 15% of the grade 3 writing samples were randomly selected to be used in this study. The grade 3 subset contained

317 out of 2,011 grade 3 writing samples, and was rated along with the rest of grade 3, using anchor papers selected from the total grade 3 writing samples. Randomly selected classrooms of grade 5 and grade 8 students also responded to the same narrative mode prompt as the grade 3 students. The across-grades set of writing samples included 180 grade 5 narrative writing samples and 172 grade 8 narrative writing samples, as well as the grade 3 subset of 317 writing samples. The across-grades set of writing samples was rated using anchor papers selected from the across-grades writing samples from grades 3, 5, and 8.

Procedure

Students responded to the writing prompt over two days in separate 50-minute periods. The two periods included a drafting session and final writing session. Teachers read aloud the instructions as they appeared in a prepared teacher's manual.

Anchor papers were chosen from the writing samples to represent score-points in a widely used six-point, six-trait rubric (Spandel, 1996). The rubric is comprised of richly defined, six-point rating scales for each of six traits of writing quality. The six writing traits in the rubric were:

1. Ideas (well-developed, clear, and complete),
2. Organization (logical order, clear introduction and ending, effective transitions),
3. Voice (commitment to topic, originality, appropriate feeling and tone),
4. Word Choice (adds interest and understanding, enhances detail),
5. Sentence Fluency (sentences flow, have varied lengths, and ease reading), and
6. Conventions (minimal errors in grammar, punctuation, spelling, and format).

Professional raters chose the anchor papers and then the final choices were reviewed and approved by school district staff. Each of the six score points, for each analytic trait, was represented by at least one anchor paper chosen from the set of writing samples to be evaluated. Several anchors were chosen for most scale points for the within-grade scoring condition, which was the operational assessment. One anchor was chosen for each scale point for the across-grades scoring condition, which was the research-only component of the assessment.

Two different sets of anchor papers were used in scoring the samples examined in this study. One set of anchor papers was selected from the entire set of grade 3 narrative writing samples and used in scoring all grade 3 papers, including the grade 3 subset. A second set of anchor papers was chosen from a combined set of papers comprised of the narrative writing samples from grades 5 and 8 and the random subset of grade 3 writing samples. The second set was used in scoring the across-grades set of papers.

Student writing samples were scored by professional raters from a commercial testing company. Raters were randomly assigned to participate in the different scoring conditions. The across-grades scoring session was assigned different raters (N = 12) than the within-grade scoring session (N = 6). In each scoring session, a minimum of two raters read and scored each paper. For any pair of score points that differed by more than one point, another rater was called on to score the paper and provide a third rating. For this study, cases that required a third rating were excluded from analysis.

Analysis

The means of the summed observed ratings were compared using a *t*-test for dependent samples, and intercorrelations among the six writing traits within and between the two anchor paper conditions (i.e., within-grade and across-grades) were examined. Inter-rater reliabilities were calculated with single-measure, one-way random effects intraclass correlation coefficients for the ratings under each anchor paper condition and between conditions. In addition, the two sets of observed ratings for the grade 3 writing samples were analyzed using the many-faceted Rasch (one-parameter logistic) model. The many-faceted Rasch model was applied because it accommodates multiple facets in the analysis, so that student ability can be estimated while accounting for rater severity and analytic-trait difficulty. The many-faceted Rasch model (Linacre, 1994) is an extension of Rasch ordered-category and partial credit models (Andrich, 1978; Masters, 1982; Rasch, 1960/1980; Wright & Masters, 1982) and its use has been demonstrated previously in analyzing assessments of writing (e.g., Engelhard, 1992; Weigle, 1998). One of the models that was employed in this study is a three-facet model (student, rater, trait) that can be expressed as Equation 1:

$$\log(P_{nijk} / P_{nijk-1}) = B_n - R_i - T_j - F_{kj}, \quad (1)$$

where P_{nijk} is equal to the probability of student n being rated k on trait j by rater i, P_{nijk-1} is equal to the probability of student n being rated $k - 1$ on trait j by rater i, B_n is the writing ability of student n, R_i is the severity of rater i, T_j is the difficulty of analytic trait j, and F_{kj} is the difficulty of rating threshold k, relative to rating threshold $k - 1$, for trait j. Observed ratings are transformed into a linear logistic scale, excluding non-estimable examinees with perfect or zero scores. An advantage of applying the many-faceted Rasch model is that estimated student abilities, rater severities, and trait difficulties can be located along the logit scale and compared to each other. Conventionally, the mean of the measurement agents within a many-faceted Rasch model are constrained to zero, with the primary object of measurement estimated with respect to this origin (Linacre, 2007).

Thus, all facets other than student ability are constrained to have their mean equal to zero in this study.

The three-facet model (i.e., students, raters, traits) was applied to the within-grade ratings and again to the across-grades ratings. Observed ratings and Rasch student-ability estimates from each anchor paper condition analysis were compared using *t*-tests for dependent samples. Patterns among the rater-severity estimates and trait-difficulty estimates within each anchor paper condition were examined as well. Raters were expected to differ in level of severity within each anchor paper condition. Rater-severity estimates were examined to see if the pattern of severity was substantially different between the anchor paper scoring conditions. Trait-difficulty estimates were also examined to find out whether analytic traits were distributed similarly for each anchor paper condition. Additional statistics that accompany the many-faceted Rasch model analysis include infit and outfit mean-square statistics, a separation index, and a reliability of separation index. Infit is a weighted mean-square residual that is influenced by unexpected observations near the estimated parameter level (of the rater or trait) and outfit is an unweighted mean-square residual that is sensitive to unexpected extreme observations and outliers. No strict guidelines for interpretation currently exist, but many researchers look for values between 0.5 and 1.5, with 1.0 indicating best fit. Linacre and Wright (1994) suggested a range of 0.4 to 1.2 as reasonable for tests that involve judgments. The separation index is a ratio of the standard deviation of the parameter estimates, adjusted for measurement error, to the root mean-square error (RMSE). Lower values of the separation index (i.e., near 1.0) indicate the extent to which raters are equally severe, with higher values indicating increasing variance among raters. The reliability of separation, a ratio of "true" variance to observed variance, provides a measure of the extent to which the parameter estimates can be reliably distinguished from each other. The reliability of separation, like Cronbach's alpha, should be high for examinees and items, or in this case, for traits. However, for raters, a higher reliability of separation indicates more distinction among raters (i.e., a lower degree of consistency).

A four-facet model (students, anchor paper conditions, raters, and traits) was also applied to the grade 3 subset of combined within-grade and across-grades ratings, in order to estimate the degree of difference between the two anchor paper conditions. The four-facet model is expressed in Equation 2 as follows:

$$\log(P_{nmijk} / P_{nmijk-1}) = B_n - C_m - R_i - T_j - F_{kj}, \quad (2)$$

where the additional facet, C_m, is the degree of challenge associated with anchor paper scoring condition *m*.

The lack of connectedness among raters between the two conditions precluded a conventional analysis to investigate whether the raters differed between groups.

To compare the disjoint subsets, the four-facet model was applied, holding the effect of rater group constant between anchor paper conditions. Rater severity estimates were group-anchored at no mean difference to allow the maximum degree of difference between the level of challenge for each scoring condition to be estimated. Raters were expected to differ in level of severity within each anchor paper condition.

To illustrate the potential impact of anchor paper selection on the assessed quality of student writing, the ratings sets were compared against the district performance standard for grade 3. A contingency table is provided to show the classifications of students (i.e., at or above standard or below standard) based on the two sets of ratings for the same papers. Cohen's kappa was computed to assess the extent of agreement between the within-grade and across-grade rating contexts with respect to student proficiency classification.

RESULTS

Observed Ratings

The grade 3 ratings of the same essays differed in magnitude and rank order when scored against different sets of anchor papers. Observed ratings were higher for the papers rated against the within-grade anchor papers. The mean summed score-points were 20.7 (SD = 3.76) and 17.0 (SD = 4.32), for the within-grade and across-grades ratings, respectively, with $t(316) = 23.47$, $p < .01$. The 95% confidence interval for the difference extends from 3.40 to 4.02 points. The correlation between raw scores from the two anchor paper conditions was .73, accounting for just over half of the variance between the two sets of ratings.

As shown in Table 1, intercorrelations among observed ratings on the six writing traits were moderately high for the within-grade scoring condition (.70 to .89) and the across-grades scoring condition (.75 to .88). Intercorrelations among the six trait ratings between the within-grade and across-grades scoring conditions were lower, ranging from .52 to .74. Intraclass correlation coefficients between rater pairs were also higher for the within-grade (.56 to .66) and across-grades conditions (.66 to .71), than between the within and across conditions (.17 to .43).

Student Ability Parameter Estimates

The three-facet analyses produced Rasch student-ability parameter estimates (expressed in logits) that were also significantly higher ($M = –2.57$, $SD = 3.88$) for the within-grade anchor paper condition than the across-grades anchor paper condition ($M = –3.84$, $SD = 3.52$), with $t(316) = 8.77$, $p < .01$. The 95% confidence interval for the mean difference extends from 0.98 to 1.55 logits. The correlation

TABLE 1
Correlations Among Observed Ratings on Six Writing Traits for the Within-Grade Scoring Condition and Between the Within-Grade and Across-Grades Conditions (bold font)

	Within-Grade Scoring Condition						*Across-Grades Scoring Condition*					
	I	*O*	*V*	*WC*	*SF*	*C*	*I*	*O*	*V*	*WC*	*SF*	*C*
W: Ideas	—	.89	.82	.82	.77	.74	**.63**	**.58**	**.61**	**.57**	**.57**	**.55**
W: Organ		—	.80	.82	.76	.75	**.63**	**.61**	**.62**	**.59**	**.58**	**.58**
W: Voice			—	.79	.72	.70	**.59**	**.57**	**.61**	**.54**	**.52**	**.52**
W: WordC				—	.80	.79	**.63**	**.61**	**.60**	**.61**	**.59**	**.58**
W: SentF					—	.86	**.63**	**.66**	**.64**	**.67**	**.66**	**.65**
W: Conv						—	**.66**	**.70**	**.67**	**.69**	**.68**	**.74**
A: Ideas							—	.81	.86	.83	.78	.76
A: Organ								—	.83	.86	.86	.84
A: Voice									—	.83	.80	.75
A: WordC										—	.88	.83
A: SentF											—	.87
A: Conv												—

Note. W = Within-grade, A = Across-grades, I = Ideas, O = Organization, V = Voice, WC = Word Choice, SF = Sentence Fluency, C = Conventions. Correlations are between the means of each pair of observed ratings for each condition, that is, under each condition, each writing sample received two ratings (two different raters) for each trait.

Intraclass correlation coefficients between raters obtained by trait, for the Within condition, Across condition, and between the Within and Across conditions were:

Within (between first and second raters): I = .59, O = .56, V = .62, WC = .63, SF = .65, C = .66;
Across (between first and second raters): I = .68, O = .67, V = .66, WC = .68, SF = .71, C = .70;
W by A (between mean W and mean A): I = .42, O = .17, V = .47, WC = .28, SF = .42, C = .43.

between ability estimates from the two anchor paper conditions was .76, accounting for under 60% of the variance between the conditions.

Analytic Trait Difficulty Parameter Estimates

The relative difficulty of the six analytic traits differed depending on whether the writing samples were scored against the within-grade anchor papers or the across-grades anchor papers. Also, the range of difficulty was more restricted for the within-grade anchor paper condition, with trait-difficulty estimates ranging from –1.01 for the least difficult trait of Word Choice to +.96 for the most challenging trait of Conventions. The range of difficulty for the across-grades anchor paper condition extends from –1.82 for Voice to +2.36, for Conventions. Consequently, the trait locations were more variable under the across-grades condition (M = .00; SD = 1.30) than for the within-grade condition (M = .00; SD = 0.62). Trait-difficulty estimates under the different anchor paper conditions are most different

TABLE 2
Three-Facet Analyses: Trait-Difficulty Estimates in Logits, with Standard Errors and Mean-Square Fit Statistics for Each Analytic Trait by Anchor Paper Condition

	Within-Grade Condition[a]			*Across-Grades Condition*[b]			*Difference*
Trait	*Logit (SE)*	*Infit MnSq*	*Outfit MnSq*	*Logit (SE)*	*Infit MnSq*	*Outfit MnSq*	*Within–Across*
Ideas	.00 (.10)	0.9	0.7	–.95 (.07)	1.0	1.0	.95
Organization	–.01 (.10)	1.0	0.9	.38 (.06)	1.0	1.0	–.39
Voice	–.40 (.10)	1.2	1.3	–1.82 (.06)	1.1	1.1	1.42
Word Choice	–1.01 (.10)	1.0	0.9	–.30 (.07)	0.9	0.8	–.71
Sentence Fluency	.45 (.10)	1.0	0.9	.32 (.06)	0.9	0.9	.13
Conventions	.96 (.09)	1.0	1.0	2.36 (.06)	1.1	1.2	–1.40
Mean	.00 (.10)	1.0	1.0	.00 (.06)	1.0	1.0	.00
SD	.62 (.00)	0.1	0.2	1.30 (.00)	0.1	0.1	–.68

[a]Within: RMSE = .10; Adj. *SD* = .61; Separation = 6.28; Reliability = .98.
[b]Across: RMSE = .06; Adj. *SD* = 1.30; Separation = 20.19; Reliability = 1.00.

for Voice and Conventions (with differences of 1.42 and –1.40, respectively). Table 2 shows the trait-difficulty estimate locations, along with their differences (Within–Across).

For each anchor paper condition, most analytic traits differed significantly among themselves, with significant fixed chi-square values for the trait facet in both analyses, $\chi^2(5, N = 6) = 244.7$, $p < .01$ for the within-grade condition and $\chi^2(5, N = 6) = 2455.9$, $p < 01$ for the across-grades condition. The trait separation index (the ratio of the adjusted standard deviation of the trait difficulty estimates to the root mean square standard error [RMSE]) for each analysis also indicated that trait variance was much greater in the across-grades analysis than in the within-grade analysis, with separation values of 6.28 and 20.19 for the within-grade and across-grades analyses, respectively. Infit and outfit mean-square statistics ranged from 0.7 to 1.3 for the within-grade condition and from 0.8 to 1.2 for the across-grades condition.

Rater Severity Parameter Estimates

Within each three-facet analysis, raters differed significantly from each other with respect to severity, with significant fixed chi-square values for the rater facet, with $\chi^2(5, N = 6) = 236.2$, $p < .01$, for the within-grade ratings, and $\chi^2(11, N = 12) = 892.5$, $p < .01$, for the across-grades condition. Rater separation values were 6.35 and 6.10 for the within-grade and across-grades analyses, respectively. Infit and outfit values ranged from 0.6 to 1.3 in each analysis. Table 3 shows individual

TABLE 3
Three-Facet and Four-Facet Analyses: Rater-Severity Estimates in Logits, with Model Standard Errors and Mean-Square Fit Statistics

	Three-Facet Analyses						*Four-Facet Analysis*		
	Within-Grade Condition[b]			*Across-Grades Condition*[c]			*Combined*[d]		
Rater	*Logit (SE)*	*Infit MnSq*	*Outfit MnSq*	*Logit (SE)*	*Infit MnSq*	*Outfit MnSq*	*Logit (SE)*	*Infit MnSq*	*Outfit MnSq*
W1	.01 (.11)	1.2	1.3				–.33[a] (.20)	1.1	1.1
W2	–.68 (.09)	1.1	1.1				–.13[a] (.10)	1.3	1.3
W3	–.46 (.11)	1.1	1.2				–.52[a] (.13)	0.9	0.9
W4	–.41 (.09)	0.8	0.8				–.42[a] (.08)	1.1	1.1
W5	1.26 (.11)	0.7	0.6				–.31[a] (.11)	1.0	1.0
W6	.28 (.09)	1.0	1.0				–.57[a] (.10)	1.1	1.1
A1				.14 (.22)	1.2	1.1	.71 (.16)	1.0	1.0
A2				–.86 (.14)	0.8	0.8	–.42 (.08)	0.9	0.9
A3				–.51 (.08)	0.8	0.7	–1.34 (.15)	0.8	0.8
A4				.40 (.09)	1.1	1.1	1.34 (.10)	0.7	0.7
A5				–.81 (.16)	0.7	0.6	.60 (.13)	0.7	0.7
A6				.74 (.09)	0.8	0.8	1.71 (.10)	1.4	1.4
A7				.86 (.07)	1.1	1.1	1.55 (.09)	0.7	0.7
A8				1.53 (.07)	1.0	1.0	–.85 (.12)	0.9	0.9
A9				–.37 (.08)	1.2	1.3	.20 (.08)	1.1	1.1
A10				.16 (.07)	1.1	1.1	.43 (.09)	1.0	1.0
A11				–.22 (.17)	1.0	1.0	–.70 (.16)	1.0	1.0
A12				–1.06 (.11)	0.7	0.7	–.95 (.10)	0.8	0.8
Mean	.00 (.10)	1.0	1.0	.00 (.11)	1.0	0.9	.00 (.12)	1.0	1.0
SD	.65 (.01)	0.2	0.2	.75 (.05)	0.2	0.2	.86 (.03)	0.2	0.2
					Within Mean		.00	1.0	1.0
					Within *SD*		.71	0.2	0.2
					Across Mean		.00	0.9	0.9
					Across *SD*		.99	0.2	0.2

[a]Within values calibrated in the within-grade analysis and set as anchor values; across-grades values were estimated against these values.

[b]Within: RMSE = .10; Adj. *SD* = .64; Separation = 6.35; Reliability = .98.

[c]Across: RMSE = .12; Adj. *SD* = .74; Separation = 6.10; Reliability = .97.

[d]Combined: RMSE = .12; Adj. *SD* = .85; Separation = 7.06; Reliability = .98.

rater severity-estimates, means and standard deviations for the two three-facet analyses and the four-facet analysis, which is described later. The distributions of rater-severity parameter estimates along a leniency-severity continuum were not remarkably different between the two anchor paper conditions, with slightly less variability for the six raters in the within-grade anchor paper condition analyses than for the 12 raters in the across-grades anchor paper conditions.

The four-facet analysis anchored the means of the across-grades condition raters and the within-grade raters. As in the separate analyses, the combined 18 raters differed significantly among each other with respect to severity, with significant fixed chi-square values for the rater facet, with $\chi^2(17, N = 18) = 1179.7$, $p < .01$, and a rater separation index of 7.06. Infit and outfit values ranged from 0.7 to 1.4.

Anchor Paper Scoring Condition Parameter Estimates

The two anchor paper scoring conditions were compared in the four-facet anchored analysis. The degree of challenge associated with the across-grades scoring condition was 2.63 logits higher, with estimates of .52 for the within-grade condition and 3.15 for the across-grades condition. The scoring condition separation index was 37.37 (RMSE = .04, Adj. *SD* = 1.31, Reliability = 1.00).

Comparison to District Performance Standard

Table 4 displays the contingencies for classification against the grade 3 district performance standard for the two scoring conditions, given the grade 3 raw scores in this sample. Given a compensatory standard set at an average raw score point rating of "3" across all six analytic traits, 36% percent (N = 115) of the grade 3 students would obtain inconsistent results on papers rated against different anchor papers. Most of the misclassification (>35%) would occur with students who would be considered at or above the standard when rated against the within-grade anchor papers. The value of kappa was .30 (*SE* = .04), indicating a moderately low level of agreement between conditions (.21 to .30 is considered "fair" according to Landis & Koch, 1977). Note that this value of kappa also approaches a maximum value due to the asymmetrical imbalance of the marginal

TABLE 4
Grade 3: Number of Students Meeting District Compensatory Standard of Average Raw Score-Point Rating of "3" When Scored Against Different Anchor Papers

	Across-Grades Anchor Papers		
Classification	*At or Above Standard*	*Below Standard*	*Total*
Within-Grade Anchor papers			
At or Above Standard	148	112	260
Below Standard	3	54	57
Total	151	166	317

Note. Percent observed agreement is 64%. $\kappa = .30$; *SE* = .04.

totals (see, e.g., Feinstein & Cicchetti, 1990). Less imbalance in the off-diagonal cells could result in a value as low as .20, given the same proportion of agreement between conditions.

DISCUSSION

The selection of different anchor papers from either within or across grade levels significantly affected the scoring of writing quality for the grade 3 students in this study. While results may not be surprising given the different grade-level contexts, they shed light on conventional scoring practices that define relative standards within a set of writing responses to be scored. Scored against the same six-trait, six-point analytic rubric, grade 3 narrative writing samples received higher grades when scored against anchor papers chosen from grade 3 samples than when scored against anchor papers chosen from a combined set of narrative writing samples from grades 3, 5, and 8. If results on the two sets of ratings in this writing assessment are compared to the district standard for grade 3, there is a considerable difference in perceived success, depending on the anchor papers used for scoring.

The difference in score outcomes for the same writing samples was presumed to be the result of two different scoring contexts, within-grade and across-grades, where different anchor papers were selected to represent the rubric. The effect of the different raters, a well-documented source of variance in writing assessment, was also a potential source of variance in this study. While a difference in the overall severity of the rater groups may have contributed to some of the difference between the ratings under the two scoring conditions, the effect of the anchor conditions, when accounting for the effect of raters in the four-facet analysis, was substantial. The anchor papers chosen to represent the score points in the rubric clearly reflected different interpretations, given the collection of writing samples to be scored. The selected anchor papers serve as proxies for the rubric, reflecting a specific scoring context and becoming the de facto scoring key. While transforming the scoring standards from the rubric to a specific context such as grade level or mode of writing may be customary practice in standardized assessments of performance, little attention has been given to the impact of anchor paper selection on score variance.

In practice, major discrepancies in the scoring of the same papers are resolved because they occur within the same scoring context. What is not known in any given scoring context, however, is the degree to which the application of the rubric has been adapted to grade-level expectations. Grade levels may have distinct expectations associated with them that are not captured in the six-trait rubric. If, for example, grade-level standards are used to augment the rubric in the selection of anchor papers, to what extent do grade-level expectations guide selection

compared to universal characteristics of effective writing? Low score outcomes on a grade 3 assessment may reflect low quality of writing against grade 3 expectations or developmentally appropriate quality against a broad-construct interpretation, depending on the context defined within range finding. The degree to which grade-level expectations are incorporated in range finding will affect inferences regarding progress over time. In discussing validity with respect to analyzing changes over time, Messick (1989) stated, "changes over time might be investigated in terms not only of stability or the relative ordering of individual scores but also of theoretically expected changes in score level and variability or in dimensional structure" (p. 55). In the present case of judging writing quality of the same samples, but employing different anchor papers, the difference in scores appears to reflect raters' different expectations for writing performance at different grade levels. The evidence for context-dependent variance in writing scores in this study seems to suggest that explicit incorporation of context-relevant standards in range finding, such as grade-level expectations, may be essential to the valid rating of writing, given perceived differences in the difficulty of analytic traits within different scoring contexts.

Anchors were selected to represent the entire range of points for each trait, regardless of whether they were drawn from only grade 3 or grades 3, 5, and 8, and without explicit direction to interpret the rubric via grade level. Even if contextual standards are explicitly used to guide range finding, operational standards may still be defined in a manner that depends heavily on the sample of papers from which the anchors are chosen. The degree to which rubric traits are differentially defined at different grade levels through the selection of anchor papers may be determined more by the distribution of student performance within the samples at hand than by actually aligning performance to grade-level performance definitions.

The anchor paper conditions compared in this study were dramatically different and yielded anticipated differences; however, the comparison exposes issues that have not been investigated with respect to the selection of anchor papers. In addition to concerns regarding the interpretation of progress over time, there is also the issue of reliability of anchor paper selection within the same context. Will different range finders choose the same anchor papers? Will different sets of anchor papers produce the same scoring outcome? What features of a range finding session support reliable anchor paper selection? A major concern is the relative nature of anchor selection. Will range finders assign anchors to every scale point for every analytic trait, even if the samples of writing represent restricted performance? While it is not uncommon to have fewer anchor papers chosen to represent extreme scale points (i.e., 1 and 6), it is rare to have none, or to have few or none for the range of more commonly applied scale points. Would range finders still find anchors for scale points 5 and 6 in a set of writing samples previously assessed at low levels of performance?

The relationship between the trait-difficulty locations also revealed that while the estimated difficulty of analytic traits differed considerably between the two scoring conditions, some traits may be more strongly related between the conditions than others. The trait-difficulty estimates for Word Choice, for example, were less related between the two anchor paper conditions. This suggests that some aspects of a broader construct of writing ability may be comparable across grade levels, while other aspects of writing ability may be defined and assessed very differently depending on the writer's grade level. The results suggest a need for further research regarding the perceptions of raters with respect to rubric interpretation and the construct of writing quality.

The selection of anchor papers is an essential part of the scoring process that directly affects scoring outcomes. Elements of the assessment context not explicit in the scoring rubric impact score variance via anchor selection, threatening the reliability and validity of direct writing assessment. The extent to which score variance depends on identifiable features of the scoring context (e.g., rater expectations associated with grade level of assessment) or on unknown aspects of the scoring context (e.g., the range of performance observed within a particular sample) is not known. Adapting scoring procedures explicitly to incorporate identifiable context-dependent scoring elements, such as expectations related to grade level or discourse mode, is recommended. Further research regarding the selection and use of anchor papers in scoring is needed to understand better the role of the anchor paper as a critical element in direct writing assessment. Results confirm the need for continued investigation into sources of variance in the design and development of writing assessments and suggest caution in the use and interpretation of large-scale writing assessment scores.

REFERENCES

Andrich, D. (1978). A rating formulation for ordered categories. *Psychometrika*, *43*, 357–374.

Arizona Department of Education. (2005). *Six trait rubric—Official scoring guide for AIMS*. Retrieved February 9, 2007, from http://ade.state.az.us/sbtl/6traits.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*(4), 9–12.

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

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 271–302). Washington, DC: American Council on Education.

Daiker, D. A., & Grogan, N. (1985). *The selection and use of sample papers in holistic evaluation*. Oxford, OH: Miami University. (ERIC Document Reproduction Service No. ED305391).

Driscoll, L. A. (1996). *10 steps to district performance assessment*. Memphis, TN: Memphis City Schools, Department of Research, Standards, and Accountability. (ERIC Document Reproduction Service No. ED406396).

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*, 171–191.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*, 93–112.

Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*, 56–70.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*, 543–549.

Fenton, R., Straugh, T., Stofflet, F., & Garrison, S. (2000, April). *Improving the validity and reliability of large scale writing assessment*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Freedman, S. W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English*, *15*, 245–255.

Goertz, M. E., & Duffy, M. C. (2001). *Assessment and accountability systems in the 50 states, 1999–2000* (Research Rep. No. RR-046). Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education. (ERIC Document Reproduction Service No. ED450639).

Gordon, B., Engelhard, G., Jr., Gabrielson, S., & Bernknopf, S. (1996). Conceptual issues in equating performance assessments: Lessons from writing assessment. *Journal of Research and Development in Education*, *29*, 81–88.

Hughes, D. C., & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement*, *21*, 277–281.

Huot, B. A. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*, 237–263.

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

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education*, *13*, 121–138.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2007). *A user's guide to FACETS*. Chicago: Winsteps.com.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 380. Retrieved December 21, 2007 from http://www.rasch.org/rmt/rmt83b.htm.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice*, *11*(1), 3–9, 20.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8.

Moon, T. R., & Hughes, K. R. (2002). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice*, *21*(2), 15–19.

Mullis, I. (1984). Scoring direct writing assessments: What are the alternatives? *Educational Measurement: Issues and Practice*, *3*(1), 16–18.

Myford, C. M., & Wolfe, E. W. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement*, *3*, 300–324.

Nevada Department of Education. (2000). *Writing proficiency examination guide 2000–2001*. Carson City, NV. (ERIC Document Reproduction Service No. ED447485).

Novak, J. R., Herman, J. L., & Gearhart, M. (1996). Establishing validity for performance-based assessments: An illustration for collections of writing. *Journal of Educational Research*, *89*, 220–233.

Odell, L., & Cooper, C. R. (1980). Procedures for evaluating writing: Assumptions and needed research. *College English*, *42*, 35–43.

Quellmalz, E. (1984). Toward successful large-scale writing assessments: Where are we now? Where do we go from here? *Educational Measurement: Issues and Practice*, *3*(1), 29–32, 35.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research, 1960. Expanded edition, Chicago: The University of Chicago Press, 1980.

Roid, G. H. (1994). Patterns of writing skills derived from cluster analysis of direct-writing assessments. *Applied Measurement in Education*, *7*, 159–170.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.

Spandel, V. (1996). *Seeing with new eyes: A guidebook on teaching and assessing beginning writers* (3rd ed.). Portland, OR: Northwest Regional Educational Laboratory.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*, 25–29.

Washington Office of the State Superintendent of Public Instruction. (2001). *Grade 10 anchor set annotations from the spring 2001 Washington Assessment of Student Learning in Writing (with) presentation guide for principals*. Olympia, WA. (ERIC Document Reproduction Service No. ED466812).

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263–287.

White, E. M. (1985). *Teaching and assessing writing*. San Francisco, CA: Jossey-Bass.

Wolfe, E. W., Kao, C.-W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*, 465–492.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch Measurement*. Chicago: MESA Press.

Zhang, L. (2000). *Delaware student testing program: Report on special writing study*. Dover: Delaware State Department of Education, Assessment and Accountability Branch. (ERIC Document Reproduction Service No. ED455271).