DOCUMENT RESUME

ED 481 661 TM 035 359

AUTHOR Popp, Sharon E. Osborn; Ryan, Joseph M.; Thompson, Marilyn S.; Behrens, John T.

TITLE Operationalizing the Rubric: The Effect of Benchmark Selection on the Assessed Quality of Writing.

PUB DATE 2003-04-00

NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *Benchmarking; Elementary Education; *Elementary School Students; Interrater Reliability; *Scoring Rubrics; Selection; Writing Evaluation; Writing Tests

IDENTIFIERS *Writing Samples

ABSTRACT

The purposes of this study were to investigate the role of benchmark writing samples in direct assessment of writing and to examine the consequences of differential benchmark selection with a common writing rubric. The influences of discourse and grade level were also examined within the context of differential benchmark selection. Raters scored sets of writing samples against a common writing rubric. Ratings were completed for 317 students in grade 3, 180 I grade 5, and 172 in grade 8. Twelve raters from a commercial testing company scores the across-grades writing samples and six to seven raters cored each set of within-grade writing samples. Benchmarks used in scoring were chosen from either within a single grade or from across several grades, depending on the set of writing samples to be scored. Raw ratings were analyzed using multifacet Rasch models and were compared to hypothetical performance standards. Results show that the assessed quality of writing depends on the benchmarks chosen to define the rubric, which are described in the paper as the operational definition of the scoring rubric. (Contains 13 tables, 7 figures, and 17 references.) (Author/SLD)

OPERATIONALIZING THE RUBRIC:

THE EFFECT OF BENCHMARK SELECTION ON THE

ASSESSED QUALITY OF WRITING

Sharon E. Osborn Popp
Arizona State University

Joseph M. Ryan
Arizona State University West

Marilyn S. Thompson
Arizona State University

John T. Behrens
Cisco Systems, Inc.

Paper presented at the annual meeting of the American Educational Research Organization
Chicago, IL, April, 2003

Abstract

The purpose of this study was to 1) investigate the role of benchmark writing samples in a direct assessment of writing and 2) examine the consequences of differential benchmark selection with a common writing rubric. The influences of discourse and grade level are also examined within the context of differential benchmark selection. Raters scored sets of writing samples against a common writing rubric. Benchmarks used in scoring were chosen from either within a single grade or from across several grades, depending on the set of writing samples to be scored. Raw ratings were analyzed using multi-facet Rasch models and were compared to hypothetical performance standards. Results show that the assessed quality of writing depends on the benchmarks chosen to define the rubric, which are described in the paper as the operational definition of the scoring rubric.

Operationalizing the Rubric: The Effect of Benchmark Selection on the Assessed Quality of Student Writing

Introduction

Direct assessments of writing performance are being included in more and more large-scale testing programs, including thirty-one state assessment programs in 1999-2000 (Goertz and Duffy, 2001). Assessments of writing performance often carry high-stakes consequences despite concerns regarding reliability and validity (Gordon, Engelhard, Gabrielson, and Bernknopf, 1996; Mehrens, 1992). The purpose of this study was to 1) investigate the role of benchmark writing samples in a direct assessment of writing and 2) examine the consequences of differential benchmark selection with a common writing rubric.

Benchmarks, also known as anchor papers, exemplars, or range finders, are the writing samples chosen to define levels of performance in the scoring rubric. The chosen benchmarks 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. Benchmark papers are 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 exempfly the various score points on the rubric. During rater training and actual scoring, these benchmark papers become the focus for the raters in evaluating each student writing sample. In this study, we examine the consequences of benchmark selection on the assessed quality of writing samples scored against benchmarks selected from within a single grade versus benchmarks selected from across several grades.

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 and Johnson,

1995; Messick, 1995). The particular benchmarks chosen to represent levels of performance in the rubric would appear to be highly related to score outcome. However, research regarding the role of benchmarks in scoring direct writing assessments is surprisingly limited. We sought to investigate whether, and to what extent, benchmarks influence the ratings of students' writing. Raters scored writing samples from students in Grades 3, 5, and 8 against a common rubric. The questions asked in this study were:

1. Does the rating of the same writing samples depend on whether within-grade or across-grades benchmarks are used in scoring, despite a common scoring rubric?
2. Does the rating of student writing in a single discourse mode depend on grade level?
3. Does the rating of student writing, within a grade level, depend on discourse mode?

The research design employed two types of ratings, scored against the same rubric. One type of ratings used benchmarks selected from writing produced across several grade levels. A portion of students in each of Grades 3, 5 and 8 responded to the Grade 3 Narrative mode prompt. These writing samples were scored against benchmark papers selected from the full set of Narrative response papers written by that portion of Grades 3, 5, and 8 students. The second type of ratings used benchmarks selected from writing produced within a grade level. The writing of all students in Grade 3 was scored against benchmarks selected from grade 3 papers only. Thus, for Grade 3, there is a set of papers scored against benchmarks from within Grade 3 and also scored against benchmarks selected from across Grades 3, 5, and 8. In addition, all students in Grade 5 wrote papers in the Literary Response mode and all students in Grade 8 wrote papers in the Persuasive mode. Thus, in Grades 5 and 8, there are subsets of students who wrote to the Grade 3 Narrative mode prompt and to the grade level mode of either Literary Response or Persuasive, respectively.

Research Question 1

Question 1 addresses whether the same writing samples were rated consistently between scoring sessions that employ within-grade versus across-grades benchmarks. The same Grade 3 writing samples were scored in two separate rating sessions and results were compared. Within each scoring condition, raters used a different set of benchmark writing samples. The two sets of benchmarks represented the same rubric, but one set of benchmarks was chosen from within the set of all Grade 3 papers and the second set was selected from a set of across-grades papers. The set of across-grades papers contained a random subset of all papers from Grades 3, as well as the Grade 5 and Grade 8 responses to the Grade 3 Narrative prompt.

Research Question 2

Question 2 is concerned with the effect of across-grades benchmark selection on the rating of writing samples from different grades. The students in Grades 5 and 8 that also responded to the Grade 3 Narrative prompt were expected to perform better on the Narrative writing task than on the Grade 5 Literary-Response task or the Grade 8 Persuasive task that reflect the writing curriculum at their respective grade levels. Grades 5 and 8 students were expected to have had more practice with Narrative writing, and thus receive higher ratings of writing quality.

Research Question 3

Question 3 explores the difference in assessed quality of writing samples written in different discourse modes by the same students. Research on the effect of different discourse modes on rating outcome has found that responses to Narrative writing tasks are usually rated higher than expository and Persuasive writing tasks (see Engelhard, Gordon, & Gabrielson, 1992; Kegley, 1986; Prater and Padia, 1983). Writing samples produced by the Grades 5 and 8 students

that responded to their grade-level writing task as well as the Grade 3 Narrative task, were compared. The expectation was that students would receive higher ratings for writing in the Narrative mode, than for writing in their respective grade-level mode.

Overview

Principal concerns in direct writing assessment are whether there exists a stable, unified construct of writing ability, and whether that construct can be assessed reliably to support inferences and comparisons across different situations. The writing rubric has been designed to define standards of writing performance, and implies a standards-based assessment framework. However, benchmarks used in the actual rating of writing are selected from the set of performances to be rated. The operationalization of the rubric by selecting benchmarks from a given set examinee performances suggests a relative assessment framework. We examine the implications of differential benchmark selection with respect to defining the construct of writing ability and characterizing the assessment framework that underlies direct writing assessment.

Method

Design

Raters scored writing samples produced by students in Grades 3, 5, and 8 against a common rubric. Students produced the samples of writing in response to grade-level prompts (Narrative, Literary Response, and Persuasive, respectively). Approximately 15% of students, randomly selected, in each of Grade 5 and Grade 8 were also asked to produce writing samples in response to the Grade 3 Narrative prompt. Approximately 15% of Grade 3 student writing samples were randomly selected to be re-rated. These Grade 3 papers were re-rated using benchmark papers chosen from across the three grades that responded the same Narrative prompt. Figure 1 illustrates the administration and scoring design.

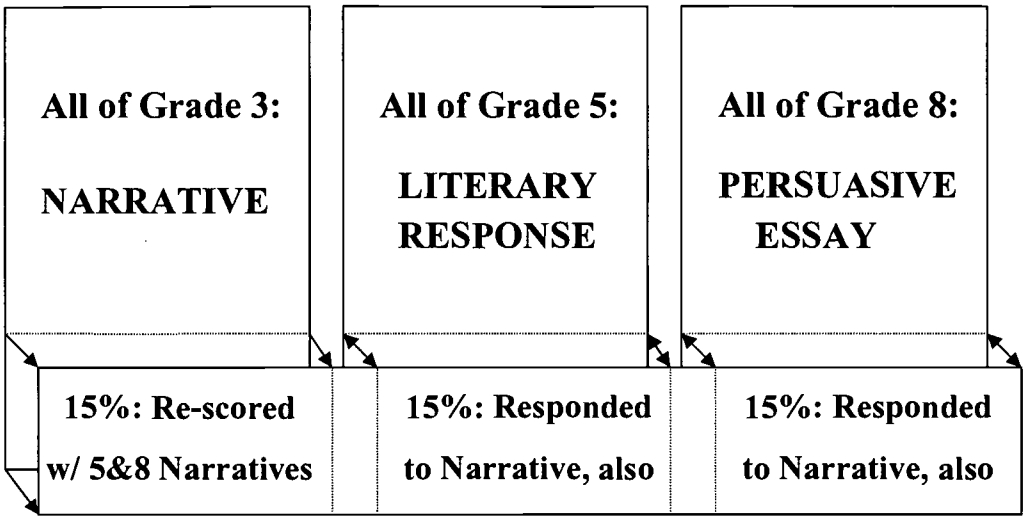

Figure 1. Administration and scoring design for the direct writing assessment.

Benchmarks to be used in scoring were chosen from each of four sets of writing samples to be scored. Grade 3 Narrative samples, Grade 5 Literary Response samples, and Grade 8 Persuasive writing samples were scored in three separate within-grade rating sessions. An across-grades rating session was also held to score the Grades 5 and 8 responses to the Grade 3 task, along with the random subset of the original Grade 3 responses.

Sample

All subjects that produced writing samples used in this study were Grades 3, 5, and 8 students from a large metropolitan school district. The assessments administered were criterion-referenced tests designed to provide direct assessment of student writing ability for students in each of the three grades. The writing assessments were part of an on-going district-wide assessment program intended to reflect curricular objectives in writing and language arts. For each grade, a writing prompt from a different discourse mode was presented: Grade 3 students responded to a Narrative mode prompt, Grade 5 students responded to a Literary Response mode

prompt, and Grade 8 students responded to a Persuasive Essay mode prompt. Randomly selected classrooms of Grade 5 and Grade 8 students also responded to the Narrative mode prompt. The set of ratings for Grade 3 contained ratings for 317 students. The Grade 5 final set contained ratings for 180 students. The Grade 8 final set contained ratings for 172 students.

Instrument

The writing performance assessment consisted of three forms. Each form contained a writing prompt written to elicit student writing in a particular discourse mode. All forms contained instructions regarding an allowed drafting period and a final writing period. Instructions also included a brief checklist that the students could use to draw their attention to the six analytic writing traits as they wrote. The form used for the within-grade assessment of Grade 5 also included a separate handout that contained the story to be read before responding to the Literary Response prompt.

The prompt for each discourse mode was chosen to align with the district curriculum and state standards at the different grade levels. Each was written to be broad enough to allow students flexibility within the limits of the prompt and to require no specific content knowledge. The prompts are shown in Figure 2.

Narrative Prompt (Within-grade prompt, Grade 3; Across-grades prompt, Grades 3, 5, and 8)

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.

Literary Response Prompt (Within-grade prompt, Grade 5)

Imagine that you are one of the characters in "The Lion and the Mouse." Which character are you, and how do you feel? Use examples from the story and tell how your feelings may have changed from the beginning to the end of the story.

Persuasive Essay Prompt (Within-grade prompt, Grade 8)

Most people have favorite entertainers, sports teams, types of pets, fast food restaurants, or places to visit. Choose your own favorite. It may be from the list above or another favorite you have. Write a paper to persuade someone else that your choice is great.

Figure 2. Writing prompts.

Administration

Students responded to the writing prompts in December, 1998. Assessments were administered over two, separate 50 minute periods. Randomly selected classrooms of Grade 5 and Grade 8 students were assessed over an additional two, separate 50 minute periods to obtain writing samples in the Narrative discourse mode. Teachers were required to read aloud the instructions as they appeared in a prepared teacher's manual.

Scoring

Professional raters from a commercial testing company scored the student writing samples in the early months of 1999. Twelve different raters scored the across-grades writing samples and six to seven raters scored each set of within-grade writing samples. In each scoring session, two different raters read and scored each paper. For any pair of score points that differed

by more than 1 point, another rater was called upon to score the paper and provide a third rating. For this study, cases that required a third rating were excluded from analysis.

Raters scored the writing samples, using a six-point, six- trait rubric (Spandel, 1996). The six writing traits evaluated 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 benchmarks and the final choices were reviewed and approved by school district staff. Each of the six score points, for each analytic trait, was represented by a benchmark paper chosen from the set of writing samples to be evaluated. Benchmark papers were chosen to guide scoring for each separate grade level. Benchmarks were also chosen from the combined Grades 3, 5, and 8 across-grades set of writing samples.

Four sets of ratings were analyzed in the study. Three sets of ratings represented the assessed writing quality of students within each of three grade levels: Grade 3, Grade 5, and Grade 8. These three ratings sets will be referred to as the Within-grade ratings sets. One set of ratings represented the assessed writing quality of a group of students from across the three grade levels. This ratings set will be referred to as the Across-grades ratings set. Only the students that had ratings in a Within-grade ratings set and the Across-grades ratings set were analyzed in the current study.

Procedure

Raw ratings were analyzed using multi-facet Rasch models. Raw ratings and Rasch-estimated student abilities, trait difficulties, and rater leniency-severity parameters were examined. The multi-facet Rasch model is an extension of the Rasch model (Rasch, 1960/1980; Wright and Stone, 1979) that accommodates multiple facets in the analysis. Student ability is estimated while accounting for rater severity and analytic-trait difficulty. The multi-facet (also called many-facet and many-faceted) Rasch model (Linacre, 1989) is an extension of Rasch ordered-category and partial credit models (Andrich, 1978; Masters, 1982; Wright and Masters, 1982) and its use has been demonstrated previously in analyzing assessments of writing (e.g., Engelhard, 1992). The multi-facet Rasch model that was employed in this study can be expressed as Equation 1,

$$\log(P_{nijk} / P_{nijk-1}) = B_n - R_i - T_j - F_k , \qquad (1)$$

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_k is the difficulty of rating threshold k, relative to rating threshold k – 1. Observed ratings are transformed into a linear logistic scale (in log-odds units, or logits) that ranges from $-\infty$ to $+\infty$. Perfect scores and zero scores are eliminated from analysis because they are non-estimable. Estimated student abilities, rater severity, and trait difficulty can be located along this scale and compared to each other. The distributions of latent trait locations within each ratings set for students, raters, and traits were examined.

Research Question 1.

The ratings from the Grade 3 Within-grade scoring were compared to the ratings from the Grade 3 subset of the Across-grades scoring. Raw ratings and Rasch parameter estimates were examined and compared between the different benchmark paper conditions. Rasch student-ability locations from each benchmark condition analysis were compared using a *t*-test for dependent samples. Patterns among the rater severities and trait difficulties within each benchmark conditions were examined, as well.

To illustrate the impact of benchmark selection on the assessed quality of student writing, the ratings sets were compared against hypothetical performance standards. Contingency tables are provided to show the classifications of students (i.e., at or above standard or below standard) based on two sets of ratings for the same papers. The proportion of misclassified students is reported for each of two hypothetical performance standards.

Research Question 2.

The rating of writing performance is compared among students from different grade levels on a common prompt in the Narrative mode. The Across-grades ratings set contains the subgroup of Grades 3, 5, and 8 students that responded to the Narrative mode prompt and were scored against benchmark papers chosen from the entire set of writing samples from the three grade levels. Raw scores were summarized for each grade level within the results for the Across-grades Grades 3, 5, and 8 ratings to describe the relative performance of students in the three grades on the writing task. An analysis of variance (ANOVA) that included an orthogonal linear contrast was conducted on the Rasch parameter estimates to assess whether the grade level means were significantly different from each other, and to test the presence of a linear trend

across the increasing grade levels. The percentage of students in each grade scoring above the median for each grade level was also reported, to provide an indication of grade-to-grade overlap. Again, relationships among rater severities and trait difficulties between the benchmark conditions are also examined and reported.

Research Question 3.

The ratings of writing by the same students were compared on different discourse mode writing tasks. The tasks were scored with the same rubric, but again, different papers represented the scoring benchmarks. For Grade 5, the ratings from the Literary Response mode were compared to the ratings from the Grade 5 subset of the Across-grades Narrative mode. For Grade 8, the ratings from the Persuasive mode were compared to the ratings from the Grade 8 subset of the Across-grades Narrative mode. Raw ratings and Rasch parameter estimates were examined and compared between the different modes for each grade level. Rasch student-ability locations from each discourse mode analysis were compared using a *t*-tests for dependent samples. Relationships among rater severities and trait difficulties between the benchmark conditions are also examined and reported. Rater severities were also compared directly for Grade 5, due to sufficient interconnectedness among the raters under both conditions, that was not present among the other rating sessions.

As with Research Question 1, the impact of writing in different discourse modes on assessed quality of student writing was examined with respect to hypothetical performance standards for the Grades 5 and 8 ratings. Contingency tables are provided to show the classifications of students based on two sets of ratings for the same students. The proportion of misclassified students is reported for each of two hypothetical performance standards, for each grade level.

Results

Research Question 1 Results

The selection of different scoring benchmarks from either within or across grade levels did affect the assessment of student writing quality for the Grade 3 students. Despite being scored against the same six-trait, six-point analytic rubric, Grade 3 Narrative writing samples received higher grades when scored against benchmark papers chosen from Grade 3 samples, than when scored against benchmark papers chosen from the combined set of samples from Grades 3, 5, and 8.

Ratings of the same essays differed in rank and magnitude when scored against different sets of benchmarks. Raw ratings (i.e., summed ratings on all traits, for both raters, for each student) were significantly higher for the papers rated against the Within-grade benchmarks, with a mean of summed score-points of 20.7 ($\underline{SD} = 3.76$), compared to 17.0 ($\underline{SD} = 4.32$) for the same papers rated against the Across-grades benchmarks, with a $\underline{t}$ ($df = 316$) of 23.474, $\underline{p} < .001$, $\alpha = .05$. The 95% confidence interval for the mean difference extends from 3.399 to 4.020 score points. The correlation between raw scores was .763 ($\underline{r}^2 = .5825$). Rasch student-ability location estimates were also significantly higher ($\underline{M} = -2.57$, $\underline{SD} = 3.88$) for the Within-grade benchmark condition than the Across-grades benchmark condition ($\underline{M} = -3.84$, $\underline{SD} = -3.52$), with a $\underline{t}$ ($df = 316$) of 8.769, $\underline{p} < .001$, $\alpha = .05$. The 95% confidence interval for the mean difference extends from 0.984 to 1.553 logit units. As with the raw scores, the rank-ordering of student-ability locations differed between the Grade 3 Within-grade and Across-grades benchmark conditions, with a correlation between estimates of .762 ($\underline{r}^2 = .5806$).

The distributions of Rasch rater-severity parameter estimates, or locations along a leniency-severity continuum (expressed in logits), were not remarkably different between the two

benchmark conditions. Rater-severity locations, intentionally centered at zero, spanned slightly less range for the six raters in the Within-grade benchmark condition (M = 0; SD = 0.708), than for the twelve raters in the Across-grades benchmark condition (M = 0; SD = 0.786).

The relative difficulty of the six analytic traits differed considerably, depending on whether the samples were scored against the Within-grade benchmarks or the Across-grades benchmarks. Table 1 provides the mean raw rating for each analytic trait, with all Within-grade means higher than Across-grades means. Table 2 shows the Rasch trait-difficulty locations (intentionally centered at zero in both analyses) estimated for the Within-grade and Across-grades ratings sets, and the difference between each estimate (Within – Across).

Table 1

Grade 3: Mean Raw Ratings for each Analytic Trait by Benchmark Type

	Benchmark Condition	
Analytic Trait	Within-grade	Across-grades
Ideas	3.53	3.02
Organization	3.49	2.68
Voice	3.58	3.14
Word Choice	3.57	2.89
Sentence Fluency	3.32	2.74
Conventions	3.19	2.51
Mean	3.45	2.83
Standard Deviation	.157	.232

Table 2

Grade 3: Trait-difficulty Locations for each Analytic Trait by Benchmark Type

	Benchmark Condition		
Analytic Trait	Within-grade (SE)	Across-grades (SE)	Within - Across
Ideas	.00 (.10)	-.95 (.07)	.95
Organization	.01 (.10)	.38 (.06)	-.37
Voice	-.40 (.10)	-1.82 (.06)	1.42
Word Choice	-1.01 (.10)	-.30 (.07)	-.71
Sentence Fluency	.45 (.10)	.32 (.06)	.13
Conventions	.96 (.09)	2.36 (.06)	-1.40
Mean	0	0	0
Standard Deviation	1.42	.68	1.05

The range of difficulty is more restricted for the Within-grade benchmark type, with location estimates ranging from -1.01, for the least difficult trait, Word Choice, to +.96, for the most challenging trait of Conventions. The range of difficulty for the Across-grades benchmark type extends from -1.82, for Voice, to +2.36, for Conventions. Consequently, the trait locations, intentionally centered at zero in both analyses, were more widely dispersed ($\underline{M} = 0$; $\underline{SD} = 1.4228$) under the Across-grades condition than the Within-grade condition ($\underline{M} = 0$; $\underline{SD} = 0.6789$). Trait-difficulty locations under the different benchmark conditions are most different for Voice and Conventions (with differences of 1.42 and -1.40, respectively). Figure 3 shows the two sets of trait-difficulty locations, mapped along the logit scale.

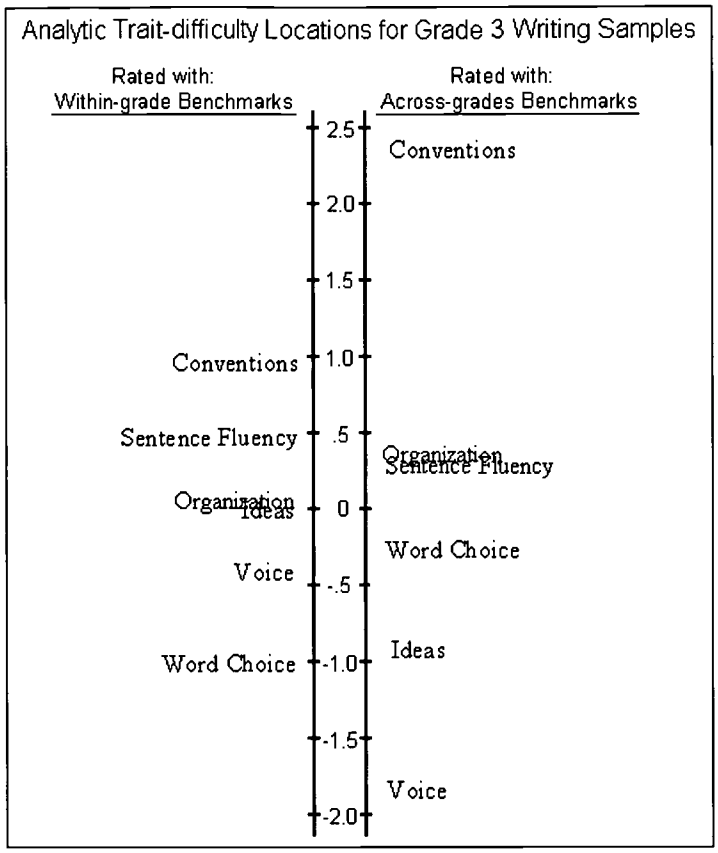

Figure 3. Trait-difficulty locations (in logit units) for Within-grade and Across-grades ratings sets for Grade 3 writing samples.

Effect on Hypothetical Performance Standards: Grade 3.

Given a compensatory standard set at an average raw score-point rating of 4 across all six analytic traits, fourteen percent of Grade 3 students would obtain inconsistent results (i.e., at or above standard under one benchmark condition and below standard on the other) on papers rated against different benchmarks. If a lower hypothetical standard is explored, such as an average raw score-point rating of 3 across analytic traits, 36% of students are classified differently between the two benchmark conditions. Most of the misclassification occurs with students who would be considered at or above the standard when rated against the Within-grade benchmarks. Forty-three percent of these students would be considered below standard when rated against

Across-grade benchmarks. Tables 3 and 4 display the contingencies for each hypothetical standard scenario, given the Grade 3 raw scores in this sample.

Table 3

Grade 3: Number of Students Meeting Hypothetical Compensatory Standard of Average Raw Score-point Rating of "4"when Scored Against Different Benchmark Papers

Classification		Across-grades Benchmarks		Total
		At or above Standard	Below Standard	
Within-grade Benchmarks	At or above standard	14	41	55
	Below Standard	3	259	262
Total		17	300	317

Table 4

Grade 3: Number of Students Meeting Hypothetical Compensatory Standard of Average Raw Score-point Rating of "3" when Scored Against Different Benchmark Papers

Classification		Across-grades Benchmarks		Total
		At or above Standard	Below Standard	
Within-grade Benchmarks	At or above standard	148	112	260
	Below Standard	3	54	57
Total		151	166	317

Research Question 2 Results

The assessed quality of writing, on the same writing task, was different for the students from different grade levels in this study. The range of raw scores overlapped considerably, but

mean raw scores, mean raw score-point ratings within each analytic trait, and mean student-ability parameter estimates differed significantly among grade levels.

Means of raw ratings increased with grade level, with a Grade 3 mean of 17.0 (SD = 4.10), a Grade 5 mean of 20.8 (SD = 4.14), and a Grade 8 mean of 24.0 (SD = 4.46). The overall mean of summed raw score points was 18.8 (SD = 5.11). The median raw score was 20.0 for the combined grades. The percentage of students in each grade that scored at or above the median was 24% for Grade 3, 64% for Grade 5, and 84% for Grade 8. Figure 4 shows the histograms of each grade level's raw scores overlaid upon each other.

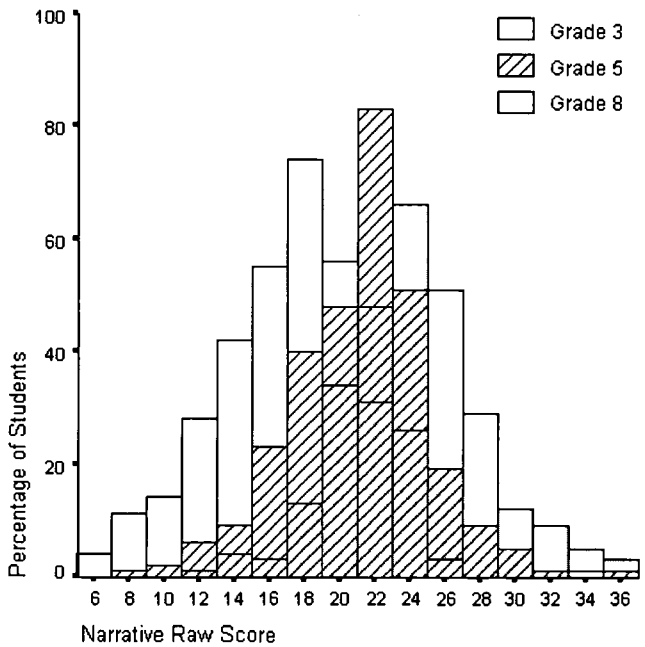

Figure 4. Overlaid histograms of Grades 3, 5, and 8 raw scores on the combined grades Narrative ratings.

Means of Rasch student-ability location estimates also increased significantly with grade level, with a Grade 3 mean of -3.84 (SD = 3.52), a Grade 5 mean of -0.13 (SD = 3.90), and a

Grade 8 mean of 3.05 (SD = 4.38). An analysis of variance (ANOVA) with an orthogonal linear contrast produced a significant F (2, 666) = 184.904, $p < .001$, $\alpha = .05$. The 95% confidence interval extends from -4.2291 to -3.4513 logit units for the Grade 3 mean, -0.7066 to 0.4412 logit units for the Grade 5 mean, and 2.3886 to 3.7067 logit units for the Grade 8 mean. The contrast estimate was significant at 4.814, with a 95% confidence interval extending from 4.221 to 5.406.

Rater-severity locations, intentionally centered at zero, ranged from -1.06 to 1.53 (M = 0; SD = 0.7861). The relative difficulty of the six analytic traits, in raw score-points, was nearly identical for the three grades, with Voice being the least difficult trait and Conventions the most difficult. Mean raw ratings for each trait are shown in Table 5. The trait-difficulty locations, also centered at zero, ranged from -1.82 to 2.26 (M = 0; SD = 1.4228). Table 6 presents the estimated trait-difficulty locations.

Table 5

Grades 3, 5, and 8: Mean Raw Ratings for each Trait, Overall and by Grade Level

	Mean Score-point			
Analytic Trait	Overall	Grade 3	Grade 5	Grade 8
Ideas	3.55	3.01	3.62	4.14
Organization	3.26	2.66	3.31	3.94
Voice	3.71	3.14	3.80	4.34
Word Choice	3.42	2.88	3.51	4.01
Sentence Fluency	3.30	2.72	3.40	3.90
Conventions	3.06	2.50	3.16	3.66
Mean	3.38	2.82	3.47	4.00
Standard Deviation	.229	.227	.228	.230

Table 6

Grades 3, 5, and 8: Trait-difficulty Locations for Across-grades Narrative Ratings

Analytic Trait	Difficulty Location (SE)
Ideas	-.95 (.07)
Organization	.38 (.06)
Voice	-1.82 (.06)
Word Choice	-.30 (.07)
Sentence Fluency	.32 (.06)
Conventions	2.36 (.06)
Mean	0
Standard Deviation	1.42

Research Question 3 Results

The assessed quality of student writing does not appear to be directly comparable on tasks that differed in discourse mode for the Grade 5 and Grade 8 student writing samples analyzed. Results did show similar measures of central tendency on student raw scores and Rasch ability locations for Grade 5, but differences for Grade 8. While strong relationships between trait-difficulties on different modes were seen for both grades, ratings for different modes did lead to inconsistencies in how students were classified based on various performance standards. Surprisingly, Grade 5 students were generally rated lower on their Narrative writing samples than on their Literary Response samples. Grade 8 students received higher ratings, in general, on their Narrative samples, than on their Persuasive samples.

Grade 5.

Raw ratings were similar for the Grade 5 Literary Response and Narrative mode papers, with mean summed score-points of 21.2 (SD = 5.33) and 20.8 (SD = 4.14), respectively, and a

paired samples $\underline{t}$ (df = 179) of 1.274, $\underline{p}$ = .204, α = .05. The 95% confidence interval for the mean difference extends from -0.238 to 1.104 score points. Rasch student-ability location estimates were not significantly different, with $\underline{M}$ = 0.20 ($\underline{SD}$ = 4.27), for the Literary Response mode ratings and $\underline{M}$ = -0.13 ($\underline{SD}$ = 3.90), and a $\underline{t}$ (df = 179) of 1.182, $\underline{p}$ = .239, α = .05. The 95% confidence interval for the mean difference extends from -0.226 to 0.900 logit units.

Raw scores were moderately correlated between modes for Grade 5, with $\underline{r}$ = .561 ($\underline{r}^2$ = .3145). Student-ability estimates were moderately correlated with each other between modes after adjusting for rater severity and trait difficulty in the Rasch analysis, with a correlation between locations of .565 ($\underline{r}^2$ = .3190).

Rater-severity locations, intentionally centered at zero, were similarly dispersed for the seven raters for both discourse modes in Grade 5, with raters from the across-grades Narrative analysis slightly more severe, on average ($\underline{M}$ = 0; $\underline{SD}$ = 0.8027 for Literary Response and $\underline{M}$ = 0, $\underline{SD}$ = 0.7168 for Grade 5 Narrative). Despite the similarities in the distributions of rater-severity locations for Grade 5, the raters were not rank-ordered similarly by severity for the different modes. The Spearman rank-order correlation between the severity locations for the two modes was $\underline{r}$ = .286. Figure 5 provides the scatterplot of the rater-severity locations for the two modes.

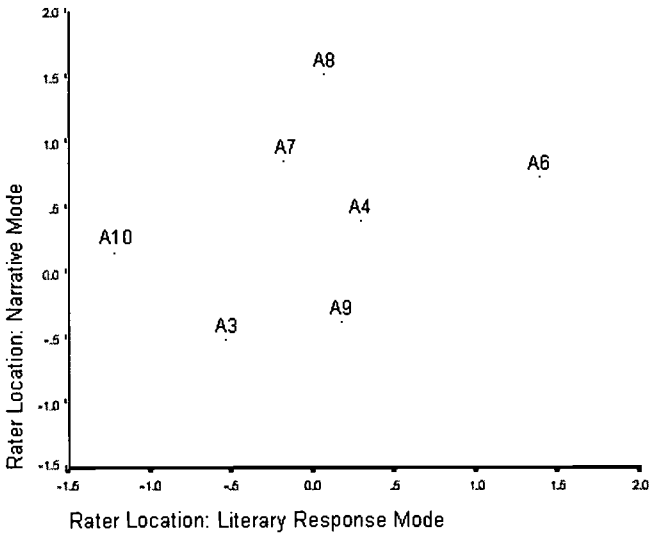

Figure 5. Grade 5 rater-severity locations: scatterplot of multi-facet Rasch model parameter estimates for Literary Response and Narrative discourse modes.

The relative difficulty of the six analytic traits differed little between modes for the Grade 5 analyses. Table 7 provides the mean raw rating for each analytic trait. Table 8 shows the Rasch trait-difficulty locations (intentionally centered at zero in both analyses) estimated for the Within-grade and Across-grades ratings sets, and the difference between modes for each trait (LitResp - Narr). The trait locations, intentionally centered at zero in both analyses, were somewhat less widely dispersed (M = 0.0017; SD = 1. 0053) in the Literary Response mode analysis than the Narrative mode analysis (M = -0.0017; SD = 1.4228). The trait-difficulty locations, shown mapped along a logit scale in Figure 6, are similar in difficulty and have nearly matching rank-orders between discourse modes (Spearman's rank order correlation of .943).

Table 7

Grade 5: Mean Raw Ratings for each Analytic Trait by Discourse Mode

Analytic Trait	Discourse Mode Literary Response	Narrative
Ideas	3.60	3.60
Organization	3.44	3.32
Voice	3.75	3.77
Word Choice	3.61	3.49
Sentence Fluency	3.43	3.40
Conventions	3.29	3.13
Mean	3.52	3.45
Standard Deviation	.164	.223

Table 8

Grade 5: Trait-difficulty Locations for each Analytic Trait by Mode

Analytic Trait	Discourse Mode Literary Response (SE)	Narrative (SE)	LitResp - Narr
Ideas	-.48 (.12)	-.95 (.07)	.47
Organization	.50 (.12)	.38 (.06)	.12
Voice	-1.26 (.12)	-1.82 (.06)	.56
Word Choice	-.70 (.12)	-.30 (.07)	-.40
Sentence Fluency	.43 (.12)	.32 (.06)	.11
Conventions	1.52 (.12)	2.36 (.06)	-.84
Mean	0	0	0
Standard Deviation	1.01	1.42	.53

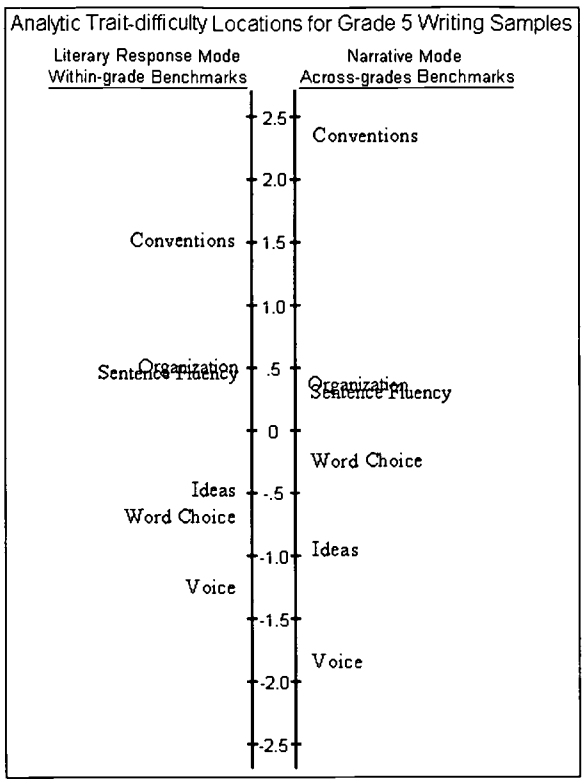

Figure 6. Trait-difficulty locations (in logit units) for Literary Response mode and Narrative mode ratings sets for Grade 5 writing samples.

Effect on Hypothetical Performance Standards: Grade 5.

Given a compensatory standard set at an average raw score-point rating of 4 across all six analytic traits, twenty-seven percent of Grade 5 students would obtain inconsistent results (i.e., at or above standard on one mode and below standard on the other) on the different discourse mode tasks. Fifty-nine percent of students who would be considered at or above the standard for Literary Response, would be classified as below standard on the Narrative task, and forty-six percent of students at or above standard on the Narrative task, would be classified as below standard on the Literary Response task. Shifting to a lower hypothetical standard, such as an average raw score-point rating of 3 across analytic traits would increase the overall percentage of students at or above standard, but produces a similar proportion of inconsistently classified

students, with twenty-four percent of students classified differently between the discourse mode tasks. Tables 9 and 10 present the contingencies for each hypothetical standard scenario, given the Grade 5 raw scores in this sample.

Table 9

Grade 5: Number of Students Meeting Hypothetical Compensatory Standard of Average Raw Score-point Rating of "4"on different Discourse Mode Tasks

Classification		Narrative		Total
		At or above Standard	Below Standard	
Literary Response	At or above standard	21	30	51
	Below Standard	18	111	129
Total		39	149	180

Table 10

Grade 5: Number of Students Meeting Hypothetical Compensatory Standard of Average Raw Score-point Rating of "3"on different Discourse Mode Tasks

Classification		Narrative		Total
		At or above Standard	Below Standard	
Literary Response	At or above standard	116	20	136
	Below Standard	23	21	44
Total		139	41	180

Grade 8.

Raw ratings were significantly different for the Grade 8 Persuasive and Narrative mode papers, with mean summed score-points of 23.0 (SD = 2.35) and 24.0 (SD = 4.46), and a t (df = 171) of –3.369, $p < .001$, $\alpha = .05$. The 95% confidence interval for the mean difference extends

from -1.552 to -0.460 score points. Rasch student-ability location estimates were also significantly different, with $\underline{M} = 1.73$ ($\underline{SD} = 2.67$), for the Persuasive mode ratings and $\underline{M} = 3.05$ ($\underline{SD} = 4.38$) for the Narrative mode ratings, with a $\underline{t}$ ($df = 171$) of -4.778, $\underline{p} < .001$, $\alpha = .05$. The 95% confidence interval for the mean difference extends from -1.860 to –0.773 logit units.

Raw scores were moderately correlated between modes for Grade 8, with $\underline{r} = .587$ ($\underline{r}^2 = .3441$). Student-ability estimates were moderately correlated with each other between modes after adjusting for rater severity and trait difficulty in the Rasch analysis, with a correlation between locations of .567 ($\underline{r}^2 = .3215$).

The distributions of rater-severity locations were not remarkably different between the different Grade 8 discourse modes. Rater-severity locations, intentionally centered at zero, were similarly dispersed for the seven raters in the Persuasive mode analysis and the seven applicable raters from the across-grades Narrative mode analysis, with $\underline{M} = 0$, $\underline{SD} = 0.7096$ for Persuasive and $\underline{M} = 0$, $\underline{SD} = 0.7168$ for Grade 8 Narrative.

The mean raw ratings for each analytic trait, provided in Table 11, appeared to be similar for the two discourse modes, with means slightly lower for the Persuasive mode than the Narrative mode. Trait-difficulty locations, estimated in the multi-facet Rasch model analyses for the Grade 8 Persuasive mode ratings and the Across-grades Narrative mode ratings, are presented in Table 12, along with their differences (Pers - Narr). The trait locations, intentionally centered at zero in both analyses, were similarly dispersed ($\underline{M} = 0$; $\underline{SD} = 1.5554$) in the Persuasive mode analysis and the Narrative mode analysis ($\underline{M} = 0$; $\underline{SD} = 1.4228$). Trait locations were rank ordered the same between the two discourse modes and are shown, mapped to the logit scale, in Figure 7.

Table 11

Grade 8: Mean Raw Ratings for each Analytic Trait by Discourse Mode

	Discourse Mode	
Analytic Trait	Persuasive	Narrative
Ideas	4.00	4.13
Organization	3.74	3.97
Voice	4.16	4.30
Word Choice	3.75	4.00
Sentence Fluency	3.77	3.96
Conventions	3.61	3.68
Mean	3.84	4.01
Standard Deviation	.202	.206

Table 12

Grade 8: Trait-difficulty Locations for each Analytic Trait by Mode

	Discourse Mode		
Analytic Trait	Persuasive (SE)	Narrative (SE)	Pers - Narr
Ideas	-1.20 (.15)	-.95 (.07)	-.25
Organization	.99 (.15)	.38 (.06)	.61
Voice	-2.39 (.15)	-1.82 (.06)	-.57
Word Choice	.03 (.14)	-.30 (.07)	.33
Sentence Fluency	.71 (.13)	.32 (.06)	.39
Conventions	1.86 (.13)	2.36 (.06)	-.50
Mean	0	0	0
Standard Deviation	1.56	1.42	.504

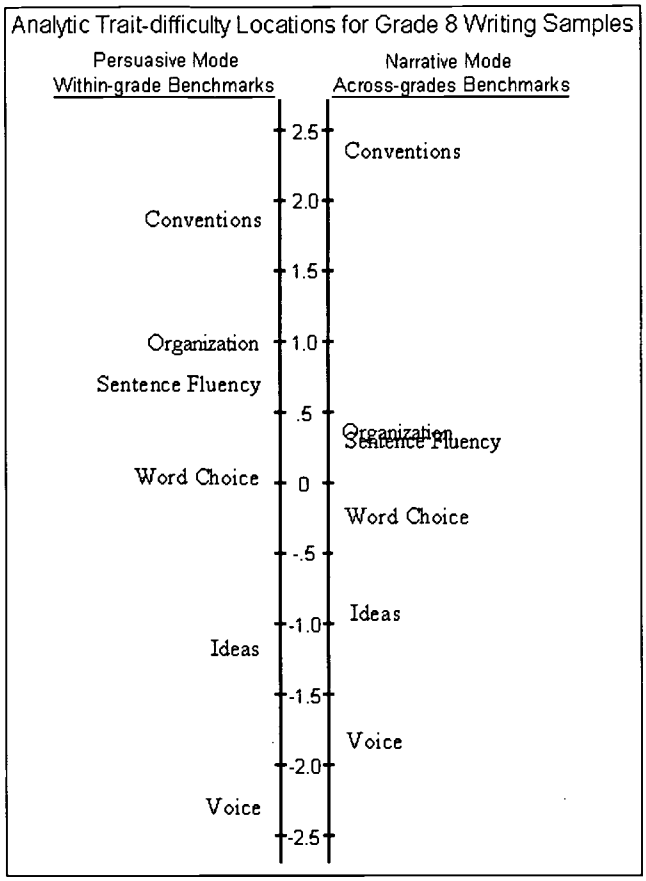

Figure 7. Trait-difficulty locations (in logit units) for Persuasive mode and Narrative mode ratings sets for Grade 8 writing samples.

Effect on Hypothetical Performance Standards: Grade 8.

Given a compensatory standard set at an average raw score-point rating of 4 across all six analytic traits, 34% of Grade 8 students would obtain inconsistent results (i.e., at or above standard on one mode and below standard on the other) on the different discourse mode tasks. A lower hypothetical standard, such as an average raw score-point rating of 3 across analytic traits would produce misclassifications of only 6.4% of students, but the overall percentage of students at or above standard, increased to 92%. Tables 13 and 14 present the contingencies for each hypothetical standard scenario, given the Grade 8 raw scores in this sample.

Table 13

Grade 8: Number of Students Meeting Hypothetical Compensatory Standard of Average Raw Score-point Rating of "4"on different Discourse Mode Tasks

Classification		Narrative		Total
		At or above Standard	Below Standard	
Persuasive	At or above standard	51	19	70
	Below Standard	40	52	102
Total		91	81	172

Table 14

Grade 8: Number of Students Meeting Hypothetical Compensatory Standard of Average Raw Score-point Rating of "3"on different Discourse Mode Tasks

Classification		Narrative		Total
		At or above Standard	Below Standard	
Persuasive	At or above standard	158	7	165
	Below Standard	4	3	7
Total		162	10	172

Discussion

Results of this study indicate that: (a) the rating of student writing depends on the benchmark papers used in scoring, (b) the rating of student writing increases with grade level, and (c) the rating of student writing may be influenced by discourse mode. Significant issues that emerged are the pivotal role of benchmark papers, whether and how specific analytic traits define the construct of writing, and whether direct writing assessment is characterized by a standards-based assessment framework or a relative assessment framework.

The same writing samples, judged against the same rubric, received different ratings when different benchmark papers were used in scoring. The selection of different scoring benchmarks from either within or across grade levels did affect the assessment of student writing quality for the Grade 3 students in this study. If results on the two sets of ratings in this writing assessment were to be compared to a hypothetical performance standard, there would be a considerable difference in perceived success, depending on the benchmarks used for scoring. Fourteen percent to 36% of students would be misclassified between the different benchmark rating conditions.

Assessed quality of writing was also found to be different for students from different grade levels in this study. Students in higher grades were expected to receive higher ratings, in general, compared to students in lower grades. As expected, mean performance on the common task increased significantly as grade level gets higher. Considerable overlap was also observed among the three grade-level ratings and results, suggesting that the assessment was not highly curriculum dependent. A high degree of grade-to-grade overlap may indicate an assessment with a low degree of curriculum dependence (Petersen, Kolen, and Hoover; 1989).

The assessed quality of student writing was not directly comparable across discourse mode for the Grade 5 and Grade 8 student writing samples in this study. Ratings across modes did share some features, such as strong relationships between trait difficulties, similar raw score-point measures of central tendency for Grade 5, and similar mean student ability locations for Grade 5. However, a substantial proportion of students were not assessed in a directly comparable manner between discourse modes for both grades. Between 24% and 27% of students would be misclassified on hypothetical performance standards between the discourse modes in Grade 5. Between 6% and 34% of students would be misclassified in Grade 8. Also

worth noting is that the relationship among the common raters that were compared between modes in Grade 5 suggested a rater by mode interaction. Rater severity estimates were ordered very differently between the two Grade 5 modes.

Students in Grades 5 and 8 had been expected to perform better on a Narrative writing task than on their respective grade-level Literary Response or Persuasive Essay tasks. Higher grade-level students would be expected to have had more practice with Narrative writing. For Grade 8, results of this study were consistent with most previous research supporting the relative difficulty of non-narrative writing. Grade 8 students received significantly higher ratings on the Narrative task than the Persuasive essay task. However, the results of the Grade 5 analyses did not show a significant difference, in general, between the ratings for the different discourse modes. Results seem to imply that writing in a grade-level discourse mode is not more challenging for Grade 5 students, compared to writing in the Narrative mode. Another interpretation may be that it is not possible to draw any conclusion regarding relative difficulty of discourse mode in this situation. Given that benchmarks were chosen from within the grade-level samples for the grade-level Literary Response mode ratings, and that benchmarks were chosen from across grades for the Grade 3 Narrative mode ratings, any difference or lack of difference might be the effect of the relative assessment framework resulting from benchmark selection. Perhaps if Grade 8 writing samples had not been included in the across-grades papers from which benchmarks were chosen, the Grade 5 students' Narrative samples might have received much higher ratings compared to their grade-level Literary Response counterparts.

Findings raise questions about the meaning and intentions underlying the rubric. The benchmarks chosen to represent the score-points in the rubric clearly reflected different interpretations, given the collection of writing samples to be scored. We might expect the writing

samples of Grade 3 students to be rated lower when compared to the performance of Grade 5 or 8 students, than when compared to the writing of same-grade peers. However, we do not expect the same writing samples of Grade 3 students, scored against the same rubric, to be rated differently. Does the rubric reflect a broad construct of writing, representing all stages of writing ability, that spans the levels of performance that extend from novice, emerging writers to expert, accomplished writers? Or is the rubric to be interpreted at varying grade levels to reflect several narrow constructs that measure writing ability relative to grade-level expectations and curricular targets? In this study, the benchmarks translated the language of the rubric into different assessments; one that seemed to measure writing at grade level and one that seemed to measure a broader construct of writing ability. The benchmarks operationalized the language of the rubric into different assessments that may reflect different contexts and perceptions of the construct of writing ability measured.

The use of uniform criteria in writing scoring rubrics clearly does not ensure consistent application of the rubric. Results of this study demonstrate that diversely defined ranges of least to highest quality could each be mapped to the generic language of a rubric. The standards of writing performance defined in the writing rubric imply a standards-based assessment framework. Benchmarks operationalize the rubric in the actual scoring of writing and are selected from the set of performances to be rated. The selection of benchmarks from a given set of examinee performances would imply a relative assessment framework. Results of this study suggest that benchmark selection does transform the standards-based assessment framework defined by the writing rubric into a relative assessment framework.

The selection of benchmarks is an instrumental part of the scoring process that directly affects scoring outcomes. Further research regarding the selection and use of benchmarks in

scoring is needed to better understand the role of the benchmark 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.

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

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. Applied Measurement in Education, 5 (3), 171-191.

Engelhard, G., Jr., Gordon, B., & Gabrielson, S. (1992). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. Research in the Teaching of English, 26 (3), 315-336.

Goertz, Margaret, E. & Duffy, Mark, C. (2001). Assessment and accountability systems in the 50 states, 1999-2000. Consortium for Policy Research in Education Research Report Series.

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 (2), 81-88.

Kegley, P. H. (1986). The effect of discourse mode on student writing performance. Educational Evaluation and Policy Analysis, 8 (2), 147-154.

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

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

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational Measurement, (3rd ed.) (pp. 221-262). New York, NY: Macmillan.

Prater, D. and Padia, W. (1983). Effect of modes of discourse on writing performance in grades four and six. Research in the Teaching of English, 17 (2), 127-134.

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.

Spandel, V. (1996). Seeing with New Eyes: A Guidebook on Teaching and Assessing Beginning Writers, 3rd ed. Portland, OR: Northwest Regional Educational Laboratory.

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

Wright, B. D., & Stone, M. H. (1979). *Best test design*: Rasch measurement. Chicago: MESA Press.

TM035359

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Operationalizing the rubric: The effect of benchmark selection on the assessed quality of student writing	
Author(s): Osborn Popp, S. E.; Ryan, J. M.; Thompson, M. S.; & Behrens, J. T.	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

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

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

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

Signature:	Printed Name/Position/Title: Sharon E. Osborn Popp, Faculty Associate/Research Associate, Arizona State University
Organization/Address: 4531 W. Toldeo Street Chandler, AZ 85226	Telephone: (480) 705-0256 FAX:
	E-Mail Address: osbornpo@asu.edu Date: 11/07/2003

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
University of Maryland, College Park
1129 Shriver Lab
College Park, MD 20742