DOCUMENT RESUME

ED 377 498 CS 214 665

AUTHOR Barrett, Thomas J.
TITLE Generalizability of Writing Tasks at Fourth Grade in the Riverside Unified School District.
PUB DATE Nov 94
NOTE 11p.; Paper presented at the Annual Meeting of the California Educational Research Association (73rd, San Diego, CA, November 17-18, 1994).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Evaluation Methods; Grade 4; Intermediate Grades; *Student Evaluation; *Testing Problems; *Test Reliability; *Writing Evaluation; Writing Research; *Writing Tests
IDENTIFIERS Riverside Unified School District CA

ABSTRACT

Students at grades four and five were administered a writing assessment that was developed to correspond to the California Learning Assessment System (CLAS) writing tasks at grade four. Teachers were trained to score the CLAS-like tasks according to the rubric developed by the State for CLAS. In addition, 164 students at three schools in the Riverside Unified School District, California, took part in the CLAS student level pilot in Spring, 1993. A generalizability study was conducted using a one facet, crossed design. The outcome was then used to conduct a decision study to determine how many tasks would be required to achieve various levels of student-level reliability. Comparisons of school level performance summaries were made between results from the CLAS assessment at grade four and results of the district's CLAS-like assessment. Results indicated that performance varied considerably both in terms of percents on the score categories as well as in mean scores. Local scorers on the CLAS-like tasks tended to place substantially more students at both extremes of the rubric than did the scorers of CLAS. In many cases, the rank ordering was markedly different. Results also indicated that at least five separate tasks scored and averaged would be required to achieve adequate levels of reliability. (Contains three tables of data.) (Author/RS)

CS

GENERALIZABILITY OF WRITING TASKS AT FOURTH GRADE IN THE RIVERSIDE UNIFIED SCHOOL DISTRICT

Prepared by:
Thomas J. Barrett, Ph.D.
Senior Program Evaluator
Educational Accountability
Riverside Unified School District

(A paper presented at the annual meeting of the California Educational Research Association Meeting, November 17-18, 1994 in San Diego, CA.)

GENERALIZABILITY OF WRITING TASKS AT FOURTH GRADE IN THE RIVERSIDE UNIFIED SCHOOL DISTRICT

Thomas J. Barrett, Ph.D.
Senior Program Evaluator

Introduction

The Riverside Unified School District has been administering a direct writing program at selected grade levels since 1991. Three writing prompts per grade level from Psychological Corporation's Language Arts Performance Assessments (LAPA) have comprised the core of the program. In addition to these assessments, students at grades four and five are administered a writing assessment that was developed to correspond to the CLAS writing tasks at grade four. Teachers were trained to score these CLAS-like tasks according to the rubric developed by the State for CLAS. Scores from this assessment are used in the district to certify whether or not students have met AB65 elementary competency standards.

Administration of these tasks at grade four in conjunction with the administration of the CLAS language arts assessment in spring of 1993 afforded the opportunity to compare performance on the CLAS with performance on tasks that were developed to coincide closely with the CLAS program. This is particularly relevant since one of the goals of SB662 was to have school districts implement CLAS compatible tasks at other grade levels. The extent to which performance on these tasks is aligned with performance on the CLAS itself will in part determine the effectiveness of a school district's efforts to assess students at various grade levels in comparable ways.

In addition, 164 students at three of our schools took part in the CLAS student level pilot in Spring, 1993. For these students, individual scores on the CLAS and the district's CLAS-like writing assessment are available. In order to determine the dependability of scores on the two tasks, a Generalizability study (G-study) was conducted using a one facet, crossed design. The outcome of this G-study was then used to conduct a Decision study (D-study) to determine how many tasks would be required to achieve various levels of student level reliability.

Method (School level comparisons)

Comparisons of school level performance summaries were made between results from the CLAS assessment at grade four and results of the district's CLAS-like assessment. These results are contained in Tables 1 and 2. It is clear from this data that performance varied considerably both in terms of percents in the score categories as well as in mean scores. Local scorers on the CLAS-like tasks tended to place substantially more students at both extremes of the rubric than did the scorers of CLAS.

Table 2 shows the rank ordering of schools based on mean scores for both CLAS and the CLAS-like assessments. In many cases the rank ordering was markedly different. The correlation between scores for the twenty-six elementary schools was .16.

Discussion

It should be noted that a number of factors could have contributed to the disparities in scores on the CLAS compared to the CLAS-like assessment at fourth grade. First, while the CLAS-like prompts were developed to mirror the CLAS as much as possible, the CLAS-like prompts were not integrated with reading/group discussion as are the CLAS writing tasks. Second, while an effort was made to train teachers to score the CLAS-like writing on a six-point rubric mirroring the CLAS scoring procedures, this training was not as extensive as that afforded the teachers scoring the CLAS writing (two hours vs. four-six hours). Also, for the CLAS-like assessment in RUSD, teachers scored their own students' papers while focusing primarily on Rhetorical Effectiveness. The Writing Conventions element was considered only when a reader was unsure about a score. CLAS papers, on the other hand, are given separate scores for Rhetorical Effectiveness and Conventions which are then weighted to arrive at a composite score.

Method (Generalizability)

A one-facet generalizability (G-study) was conducted to establish the dependability of the measurements used at fourth grade assuming the CLAS and CLAS-like tasks to be interchangeable. A randomized block design (Kirk, 1968) was used to determine the variance components of the model (Table 3). Because the variance component for items (writing tasks) was negative, it was set to zero in accordance with standard generalizability procedures.

Table 4 provides the variance component estimates and the generalizability coefficients corresponding to student assessments consisting of various numbers of tasks like the ones administered in the CLAS pilot and the district's CLAS-like program. Although G-coefficients are computed differently for relative vs. absolute decisions, the results are identical in this case because of the zero variance component for "Items". It can be seen that with one task, the G-coefficient is .50. Using a method analogous to the Spearman-Brown formula in classical test theory, this coefficient can be projected to various numbers of tasks. In order to achieve a G-coefficient of .74, three tasks would be required while five tasks would be needed to improve the dependability to .83.

Discussion

The pattern of these results is consistent with that found in other generalizability studies of direct writing although the G-coefficient of .50 was at the high end of those typically reported for single task writing samples. Since the reliability coefficient (analogous to the G-coefficient) places an upper limit on the validity coefficient it is important to know that the maximum validity obtainable if our writing tasks could be correlated to a hypothetical "true writing score" is .71. The coefficient of determination is thus .50 indicating that a maximum of 50% of the variance in our writing scores can be attributable to the underlying achievement trait when only one writing task is given to students.

Summary

To those who have studied the technical characteristics of writing assessment programs, these results should not be surprising. It is important, however, to keep reminding users of direct writing information about the dependability of information obtained from a single sample of student work. High stakes decisions should not be based on single assessments with this degree of reliability. Because of the simplicity of using a single score, school districts often rely on one writing sample to make judgments about students including the certification of AB 65 competency. Unfortunately, there continues to be little psychometric justification for doing so.

Given the reality of what it takes to achieve adequate levels of reliability in a standardized performance assessment program (probably at least five separate tasks that are scored and averaged), it is not likely that many districts will make the commitment to such an extensive, and intrusive, assessment program. Especially when writing is only one of several content areas that needs to be assessed. In all likelihood, performance

assessment will remain dichotomized between (a) single performance assessments in a couple of content areas for purposes of standardized accountability, and (b) a system of informal performance assessments such as teacher working portfolios for student level diagnostic feedback.

One of the ways that informal, classroom-based performance assessments might be used is to certify graduation proficiencies. Although this may at first glance appear to be more subjective and less defensible than a formal writing sample--given the low levels of reliability extant in single sample assessments we may be on firmer ground to rely on the expert judgment of professionals utilizing a variety of diverse, informal assessments.

Table 1
Percents in Performance Levels
CLAS vs. CLAS-Like

SCHOOL	Percents in CLAS Performance Levels					
	L1	L2	L3	L4	L5	L6
A						
CLAS	2	13	42	33	5	0
CLAS-LIKE	7	11	33	28	14	8
B						
CLAS	0	5	42	39	9	0
CLAS-LIKE	4	24	25	30	14	3
C						
CLAS	2	26	56	12	5	0
CLAS-LIKE	0	27	39	20	14	0
D						
CLAS	3	15	39	39	5	0
CLAS-LIKE	2	8	13	22	27	28
E						
CLAS	0	8	39	44	7	2
CLAS-LIKE	11	16	27	32	13	0
F						
CLAS	0	11	38	40	11	0
CLAS-LIKE	0	2	26	34	23	15
G						
CLAS	0	20	36	37	5	0
CLAS-LIKE	9	19	21	22	20	7
H						
CLAS	0	7	50	37	2	0
CLAS-LIKE	0	22	38	29	9	2
I						
CLAS	0	11	31	40	16	0
CLAS-LIKE	2	19	33	21	15	10
J						
CLAS	0	13	64	22	0	0
CLAS-LIKE	4	22	29	27	17	1
K						
CLAS	2	20	41	32	3	0
CLAS-LIKE	7	16	43	18	12	4

SCHOOL	L1	L2	L3	L4	L5	L6
L						
CLAS	0	5	48	39	7	0
CLAS-LIKE	5	21	34	25	10	5
M						
CLAS	7	14	40	29	5	0
CLAS-LIKE	2	18	27	33	16	2
N						
CLAS	7	9	35	32	11	0
CLAS-LIKE	9	29	37	18	8	0
O						
CLAS	2	2	22	51	20	0
CLAS-LIKE	6	20	25	24	18	7
P						
CLAS	0	11	49	38	0	0
CLAS-LIKE	11	17	38	21	14	0
Q						
CLAS	3	16	55	22	5	0
CLAS-LIKE	7	36	26	21	10	1
R						
CLAS	0	2	49	47	2	0
CLAS-LIKE	4	11	25	33	20	7
S						
CLAS	3	6	47	39	6	0
CLAS-LIKE	5	12	24	36	16	7
T						
CLAS	0	7	49	36	9	0
CLAS-LIKE	12	22	22	26	17	2
U						
CLAS	4	7	34	42	9	2
CLAS-LIKE	3	15	40	27	13	1
V						
CLAS	0	7	26	55	9	2
CLAS-LIKE	2	6	21	59	6	6
W						
CLAS	3	0	46	30	18	0
CLAS-LIKE	9	19	30	22	14	3

SCHOOL	L1	L2	L3	L4	L5	L6
X						
CLAS	0	15	31	36	16	2
CLAS-LIKE	8	22	29	21	13	6
Y						
CLAS	5	12	49	27	5	2
CLAS-LIKE	7	11	25	31	21	5
Z						
CLAS	2	15	50	30	2	0
CLAS-LIKE	4	17	20	32	25	1
DISTRICT						
CLAS	2	11	43	36	7	0
CLAS-LIKE	6	18	29	27	15	5

Table 2
Mean Scores and Rank Ordering on CLAS and CLAS-Like Writing
Fourth Grade

School	Mean Score CLAS	Mean Score CLAS-LK	Rank CLAS	Rank CLAS-LK
A	3.27	3.54	19	9
B	3.55	3.35	7	13.5
C	2.92	3.21	26	21
D	3.28	4.48	17	1
E	3.56	3.20	6	22.5
F	3.51	4.23	9	2
G	3.28	3.47	17	12
H	3.35	3.31	14	16
I	3.62	3.58	3.5	8
J	3.09	3.34	25	15
K	3.14	3.24	22	19
L	3.48	3.29	11	17
M	3.12	3.50	23	10
N	3.33	2.87	15	26
O	3.88	3.49	1	11
P	3.28	3.10	17	24
Q	3.10	2.94	24	25
R	3.49	3.75	10	4
S	3.39	3.67	13	5
T	3.47	3.20	12	22.5
U	3.52	3.35	8	13.5
V	3.73	3.79	2	3
W	3.62	3.23	3.5	20
X	3.59	3.27	5	18
Y	3.21	3.63	20	6
Z	3.15	3.61	21	7

Table 3
Analysis of Variance Table and Estimated Variance Components

Source of variation	Sum of Squares	df	Mean Square	Variance component
Main Effects				
Items (I)	.048	1	.048	0*
Persons (P)	309.891	164	1.89	.628
2-way interaction				
PI,e	103.952	164	.634	.634
Total	413.891	329		

* Negative value (-.0035) set equal to zero.

Table 4
Estimated Variance Components and Generalizability Coefficients for Different Decision Study Designs

	Number of Items				
Source of Variation	1	2	3	4	5
Persons (P)	.628	.628	.628	.628	.628
Items (I)	0*	0*	0*	0*	0*
PI	.634	.317	.211	.159	.127
Generalizability Coefficients					
Relative Decisions	.50	.66	.74	.79	.83
Absolute Decisions**	.50	.66	.74	.79	.83

* Negative value (-.0035) set equal to zero.
** Absolute decisions yield the same value as Relative because of zero item variance