C-R-E-S-S-T

National Center for Research on Evaluation, Standards, and Student Testing

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Interim Report: The Reliability of Vermont Portfolio Scores in the 1992-93 School Year

CSE Technical Report 370

Daniel Koretz, Stephen Klein, Daniel McCaffrey, and Brian Stecher

CRESST/RAND Institute on Education and Training

December 1993

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

INTERIM REPORT:

THE RELIABILITY OF VERMONT PORTFOLIO SCORES

IN THE 1992-93 SCHOOL YEAR[1]

Daniel Koretz, Stephen Klein, Daniel McCaffrey, and Brian Stecher
CRESST/RAND Institute on Education and Training

Summary

The 1992-93 school year saw the second statewide implementation of the Vermont assessment program. RAND, as part of the Center for Research on Evaluation, Standards, and Student Testing, continued its ongoing evaluation of the program's implementation, its effects on education, and the quality of the performance data it yields.

RAND's evaluation of the program in 1991-92 provided mixed news. On the positive side, the study found evidence that the assessment program was having a strong impact on instruction. However, the reliability of the portfolio scoring was low in both writing and mathematics—so low that most of the planned uses of the performance data were abandoned.

This interim report provides the first results of RAND's evaluation of the second year of the program, 1992-93. It discusses only the quality of the assessment data. Other important information, such as data about the program's effects on instruction, is forthcoming. This interim report is being released in advance of the full report of the RAND study because of the Vermont State Department of Education's pressing need for data to inform its decisions about the release of assessment results.

The program was altered in many ways in 1992-93, and this further evolution resulted in a clear increase in the reliability with which

---

[1] The results reported here are preliminary and have not been reviewed. Final results of RAND's evaluation of the operation of the Vermont assessment program in 1992-93 will be published by RAND later in the 1993-94 school year.

mathematics portfolios were scored. While more detailed scores are not as reliable, simple total scores for each portfolio showed moderately high reliability of scoring, particularly in the eighth grade. In contrast, the reliability of writing portfolio scores did not improve substantially and was considerably lower than in mathematics.

The Vermont State Department of Education would like to report scores for supervisory unions, local districts, or schools, and the scoring system in 1993 was designed to provide scores for supervisory unions. If total scores are used, reliability of scoring is no longer a binding constraint in mathematics for reporting average scores below the state level. However, the RAND evaluation found that supervisory union averages are not especially informative because they show very little variation and have large margins of error. The use of portfolios, rather than other types of tests, contributed relatively little to these problems. Rather, the problems appear to stem primarily from the number of students whose work is scored in each jurisdiction and the choice of supervisory unions as the reporting unit.

Finally, with the improvement of the reliability of scoring in mathematics, it is now urgent that attention be turned to other questions of validity. Reliability of scoring is necessary but insufficient to establish the validity of scores. Other questions of validity were partially moot in 1991-92 because of the low reliability of scoring, but they now should become a central focus in the program's continuing development.

Background

The Vermont assessment program was implemented statewide for the second time during the 1992-93 school year. RAND, as a part of the Center for Research on Evaluation, Standards, and Student Testing, continued its evaluation of the program in 1992-93. The RAND study, which began during the 1990-91 pilot implementation of the program, examines the implementation of the program, its effects on education, and the quality of the achievement data it produces. Because the Vermont assessment program is still in a state of development, RAND's evaluation is designed to provide periodic feedback about the strengths and weaknesses of the program to facilitate program improvement.

RAND's evaluation of the first statewide implementation in 1991-92 found mixed results. On the positive side, interviews and questionnaires administered to principals and teachers suggested that the portfolio program was having a strong effect on the instruction given to Vermont students. On the negative side, the reliability of portfolio scoring was low in both mathematics and writing. The unreliability of scoring was severe enough to preclude most planned uses of the scores.

In response to the RAND findings and comments from Vermont educators, the Vermont State Department of Education instituted a number of changes in the program for the 1992-93 school year. Perhaps most important, all portfolio scoring for state reporting was conducted at a single site during five days in June, 1993, which permitted greater standardization and control of scoring. In an effort to improve the accuracy of scoring, raters participated in calibration sessions (scoring prescored pieces and then discussing disagreements) twice a day.

What effects did these and other changes in the program have on the reliability of portfolio scores? This project memorandum presents the initial results of RAND's analysis of the reliability of the 1992-93 scores. These results are presented separately from and in advance of the rest of the results of the RAND evaluation of the 1992-93 program because of the Vermont State Department of Education's need to make decisions about the release of assessment results. Additional results—for example, further information on the effects of the program on instruction—will be presented in a later report.

Questions Addressed

This project memorandum addresses only a narrow range of questions pertaining to the quality of the assessment results. In discussing student-level scores, we focus on the reliability of ratings—that is, the extent to which raters agreed about the scores that should be assigned to portfolios or to pieces within them. This is only one aspect of the broader question of the reliability of student scores. (For example, even if rater reliability is perfect, students may show inconsistent performance over time or across alternative samples of tasks.) Although the Vermont system is not designed to produce student-level scores for use outside of schools, student scores are the basic building block of the system and constrain the quality of scores at higher levels of aggregation

(such as schools or supervisory unions). We also consider the effects of sampling of students on the reliability of supervisory union average scores.

As we discussed in earlier reports (e.g., Koretz, McCaffrey, Klein, Bell, & Stecher, 1992), rater reliability can be expressed in various ways. In this memorandum, we focus on the reliability coefficient, which is the (Spearman) correlation coefficient between the scores provided by two raters, for all portfolios that were scored twice.[2] A value of 0.00 indicates no systematic relationship between the scores provided by two raters, while a value of 1.00 indicates perfect agreement.

Reliability of Student Scores in Mathematics

The reliability of mathematics portfolio scores improved compared to 1992. The reliability of ratings for individual pieces on individual scoring dimensions remained low despite the improvement, but total scores reached a moderately high level of rater reliability.

We started by estimating the reliability of scores for individual pieces of work on single scoring dimensions. We then created a combined score (across pieces) on each of the seven dimensions and computed the reliability of those dimension-level scores. Finally, we created a single total score for each portfolio by combining across both pieces and dimensions. This total score is more reliable than the dimension-level scores because it combines information from several sources. We report the reliability of all three scores—piece-level, dimension-level, and total—here.

The reliability of single pieces on individual scoring dimensions varied among dimensions, from a low of .35 (PS4: Outcomes of activities) to a high of .60 (C2: Mathematical representation). The range of reliability coefficients was approximately the same for both grades. The average reliability coefficient for individual pieces was .50 in Grade 8 and slightly lower in Grade 4, which represents modest improvement since 1992 (Table 1).

---

[2] The number of twice-scored portfolios ranged from 155 in fourth-grade mathematics to 779 in fourth-grade writing. However, some of these portfolios could not be used in the analyses reported here because they included too few pieces (or too few that were considered scorable by both raters). The smallest number of cases underlying the reliability analyses here was 116 portfolios in eighth-grade mathematics.

Table 1

Piece-Level Correlations Between Readers, Mathematics (within-dimension correlations averaged across dimensions)

	1992	1993
Grade 4	.34	.46
Grade 8	.37	.50

The Vermont program is designed to provide composite scores for each dimension, combining across all pieces (but not across dimensions). These composite dimension-level scores are somewhat more reliable than scores on individual pieces by virtue of combining multiple sources of information. We created composites simply by combining the dimension scores across pieces.[3] The rater reliability of these dimension-level composite scores increased appreciably from 1992, reaching a high of .65 in Grade 8 (Table 2).

The reliability of total scores (created by averaging scores across pieces and dimensions) was fairly high, especially in Grade 8. Interrater correlations of .80 or higher are often considered reasonably strong for performance assessments, and the eighth-grade results nearly reached this level (Table 3). The improvement in reliability was also particularly marked in Grade 8.

Table 2

Average Correlations Between Readers, Dimension-Level Composite Scores, Mathematics

	1992	1993
Grade 4	.42	.57
Grade 8	.38	.65

***Note*. Dimension-level scores were created by averaging scores across pieces for each dimension. These composite scores were then correlated across raters. The resulting correlations were averaged across dimensions to obtain the values in this table.**

[3] In 1992, the Vermont State Department of Education used an alternative approach for creating dimension-level composite scores. We do not know what procedures the Department will follow this year.

Table 3

Correlations Between Readers, Mathematics Total Scores, Combining All Dimensions and Pieces

	1992	1993
Grade 4	.60	.72
Grade 8	.53	.79

Reliability of Student Scores in Writing

The reliability of ratings in writing was low in 1992 and—in contrast to mathematics—improved only slightly in 1993.

We began by computing the rater reliability of the scores assigned on single dimensions to individual parts of the portfolio, the best piece and the "rest." Because the parts of the writing portfolio, unlike the pieces in the mathematics portfolio, are qualitatively different from each other, we estimated reliability separately for each part. These reliabilities barely increased relative to 1992. As in 1992, the 1993 reliabilities were very similar for the best piece and the rest, and they varied only modestly from one scoring dimension to the next. Accordingly, for simplicity, we report here only a single average reliability (averaging the reliability coefficients across all five dimensions) for the best piece for each grade. In both grades, both of these reliability coefficients were low. The higher reliability was only .45 (in Grade 8), which represents a gain of only .03 from 1992 (Table 4).

More reliable scores can be obtained for portfolios by creating a single total score, combining scores across the two parts and the five scoring dimensions. The reliability of this total score increased only slightly in Grade 4 and trivially in Grade 8 relative to 1992 (Table 5).

Quality of Supervisory Union Scores in Mathematics

The intent of the Vermont program has been to provide performance information for aggregates (schools, districts, or supervisory unions) rather than for individual students. Because of the unreliability of portfolio scoring, we recommended limiting the reporting of 1992 assessment results to a

Table 4

Piece-Level Correlations Between Readers, Writing Best Pieces (within-dimension correlations averaged across dimensions)

	1992	1993
Grade 4	.35	.40
Grade 8	.42	.45

Table 5

Correlations Between Readers, Writing Total Scores, Combining All Dimensions and Both Parts

	1992	1993
Grade 4	.49	.56
Grade 8	.60	.63

statewide average score. Because the reliability of writing scores is basically unchanged, we make the same recommendation for 1993 writing scores.

Because of the appreciable improvement in the reliability of the mathematics portfolio scoring, however, it is now reasonable to consider reporting mathematics scores at lower levels of aggregation than the state as a whole. The state's scoring procedures called for scoring a sample of 30 mathematics portfolios per grade from each supervisory union. The number of portfolios scored from individual schools or local education agencies was in many cases too small to produce reliable averages for those smaller units.[4] Accordingly, the Vermont State Department of Education planned not to report any scores below the level of supervisory unions, and we have limited our analysis to date to average scores for supervisory unions. We also restricted our analysis to total portfolio scores (rather than piece-level or dimension-level scores) because they are the most reliable of the scores we computed.

[4] This is primarily because of the small number of student scores, and the problem would be similar even if a highly reliable test were used.

It is important to note that supervisory unions were selected as the unit for reporting in 1993 only as a practical compromise. To have obtained scores at a lower level of aggregation—districts or schools—would have required a larger volume of scoring than the state could afford. The Department's position was that it was important to push for reporting below the level of the state as a whole, even if the interim reporting level was not ideal.

With the improvement in scoring shown in 1992-93, the reliability of ratings is no longer a binding constraint for reporting average total mathematics portfolio scores for supervisory unions. However, for several reasons unrelated to the reliability of scoring, we conclude that supervisory union average scores are not an especially useful measure.

First, supervisory union average scores show so little variation that most of the differences among them are not informative. In Grade 4 mathematics, for example, the average scores for the 58 supervisory unions for which we had sufficient data ranged from 1.8 to 2.6 (on a scale of 1 to 4), but only a handful had scores near the ends of that range. Fifty four (93%) of the supervisory union averages fell within a range of only 0.3 point, from 2.0 to 2.3.[5] Indeed, fully 83% of the supervisory unions (48) had average scores that fell within 0.1 point of the state average of 2.2! This can be seen graphically in Figure 1; the dotted horizontal line stretching across the graph is the state average of about 2.2, and the short horizontal lines represent the mean for each supervisory union. (The vertical gray bars are confidence bands and are discussed below.)

Supervisory union averages varied more in Grade 8, but even there most supervisory unions fell within a very narrow band. In Grade 8, the state average was about 2.3, and supervisory union average scores ranged from 1.9 to 2.7. All but eight of the 59 supervisory unions for which we had Grade 8 scores (86%) had averages within the range of 2.0 to 2.5 (Figure 2).

A preliminary analysis suggests this lack of variation stemmed from the choice of supervisory unions as the reporting unit rather than a lack of variation in the quality of portfolios. In Grade 4, for example, only 5% of the total variance in portfolio scores was variation among supervisory unions. By contrast, about a fourth of the total variance in scores was attributable to

---

[5] In all comparisons in this and the following paragraph, scores and differences are rounded to one decimal place.

schools within supervisory unions. (The remainder—about 70% of the variance—was differences among students and various types of error.) While this analysis was only preliminary, it suggests that reporting at the level of schools or perhaps districts would produce a substantially greater variation in average scores. On the other hand, reporting scores for schools or districts would require additional resources and would be hindered by the very small enrollments in some schools and districts, particularly in Grade 4.

This lack of variation among supervisory unions might be expected if supervisory unions are only units of administrative convenience or geographic proximity and are not the level of governance at which critical decisions about instructional quality are made. If that is the case, educational decisions that are likely to affect scores, such as teacher selection and decisions about curriculum and instruction, are likely to be made not at the supervisory union level but rather at the level of schools or districts. If this model is correct, supervisory unions resemble somewhat random samples of the state's schools, and the average scores for such samples would be expected to cluster around the overall state average.

The second reason why supervisory union averages for 1993 do not seem useful is that they have large margins of error. The average score for a unit that includes only a small group of students can be expected to vary markedly from year to year.[6] This annual fluctuation would be expected of scores on any achievement test, not just portfolios. One of the several reasons why this occurs is the essentially random differences between successive cohorts of students. Some cohorts are particularly able or well-behaved, for example, while others include less able or more disruptive students. In addition, students perform better on some occasions than others. When the number of scores is as small as it was for supervisory unions in 1993, those fluctuations tend to be large.

We can only approximate the margins of error for supervisory union means at this time because of the limited amount of data available. Accordingly, we generated two estimates for each grade. The first estimates, called "simple random sample" estimates, are indicated by the gray vertical

---

[6] In the 1993 scoring, scoring was restricted to 30 portfolios from each supervisory union in each subject and grade regardless of the supervisory unions' enrollments because of limited resources for scoring.

bars in Figures 1 and 2. These are relatively narrow confidence bands that do not take into account the substantial clustering of portfolio scores in individual schools. (That is, scores of students selected from within a school are likely to be more similar than scores of students from different schools, all other things being equal, because students within a school share curricula, teachers, and other important aspects of schooling.) A second, much wider set of confidence bands that attempts to take these factors into account, labeled "clustered sampling assumptions," appears in Figures 3 and 4. We expect that data from additional years would show confidence bands somewhere between these two sets in size. In both cases, the bands shown in the figures are 95% confidence bands; that is, given the observed average, the probability is 95% that the "true" average falls within the band.

Even the overly narrow, "simple random sample" confidence bands show that most of the supervisory union averages from this year were statistically indistinguishable from the state average. (That is, in most cases, one cannot have reasonable confidence that differences from the state average reflect anything other than random chance.) In fourth grade, 11 supervisory union averages were lower than the state average by a statistically significant amount, and 7 supervisory unions were significantly higher than the state average (Figure 1). The other 40 supervisory unions were, statistically speaking, simply average. Using the more conservative "clustered sampling assumptions" bands, we estimate that only 4 supervisory unions were significantly lower than the state average, and only a single supervisory union was significantly above average (Figure 3).

The pattern was similar in Grade 8. For example, using clustered sampling assumptions, we estimated that only a single supervisory union average was significantly below the state average in Grade 8, and only 2 supervisory unions were significantly above average (Figure 4).

Finally, if portfolio scores are to be used to draw meaningful comparisons among units (whether supervisory unions, districts, schools, or even classes), it is essential to address many issues bearing on the validity of the scores in addition to the reliability of scoring (see Koretz, Stecher, Klein, McCaffrey, & Deibert, 1993). That is, reliability of scoring is the essential first step toward validity, but it is only the first step; even perfectly reliable scoring is not a

guarantee that scores can support the conclusions that people want to base on them.

Our efforts to assess validity for the 1991-92 program were hindered by a variety of factors, including the low reliability of scores (which precluded many of the appropriate analyses of validity), other limitations of the data, and insufficiently precise definitions of the domains of achievement that the portfolios should measure. However, our analyses revealed a number of potential threats to the validity of comparisons among schools, districts, or supervisory unions. For example, teachers had different rules about authorship and selection of tasks; they also reported substantially different amounts of revision of students' work and different guidelines about the amount of help students could receive from others. Such differences in practice could exaggerate or mask true differences in student performance.

Until this year, matters of validity—including further consideration of factors such as the variations in practice we observed—were a secondary concern, because the unreliable scoring in 1991-92 made them largely moot. However, with the improvement in the reliability of mathematics scoring, validity must now become a central focus in the ongoing development of the portfolio program.

Conclusions and Recommendations

This past year, in contrast to 1991-92, mathematics and writing presented substantially different pictures.

The progress shown in scoring mathematics portfolios is encouraging. However, the reliability of scoring, particularly below the level of total scores, still leaves room for improvement. The meaningfulness and utility of dimension-level scores, which are the intended product of the assessment program, will be severely constrained unless reliability is increased further. Additional training may improve the reliability of ratings, but other changes may be needed as well. It may be necessary to refine the scoring rubrics or to simplify them, perhaps using only one or two scoring dimensions each for problem solving (presently four dimensions) and communication (now three dimensions). It may also prove necessary to place further restrictions on the

types of tasks considered acceptable for inclusion in portfolios to permit more consistent application of the scoring criteria.

The increasing reliability of scoring in mathematics also underscores the importance of addressing questions of validity. In our opinion, some of the needed steps are conceptual rather than technical. We believe that it is essential to further clarify what domains of mathematical performance the portfolio assessment is intended to assess. It is important to set clear policies about the types of tasks that are permissible for portfolios and the types of preparation and revisions that are acceptable. We also believe it is essential to address differences in the difficulty of the tasks assigned by teachers, because these differences are currently confounded with students' scores. Clarifications of this sort will increase the validity of comparisons and will also make it more feasible to obtain the evidence needed to document validity. Validity is a continuum, however, and improvements in validity can be expected to be only incremental.

Our conclusions about the writing portfolio assessment are more pessimistic. In our opinion, it is unrealistic to expect a substantial rate of improvement in the reliability of the writing portfolio scores unless the program is changed fundamentally. Writing is in many ways a more tractable subject than mathematics for this type of assessment, and there is limited evidence from other programs that reliable scoring of writing portfolios is practical. Thus, the fact that the reliability of scoring in writing lags well behind that in mathematics strongly suggests serious flaws in the writing program. Moreover, a variety of evidence, including patterns in the writing portfolio scores, observations of the benchmarking sessions in the June 1993 scoring, and our informal analysis of the writing rubrics, is consistent with the hypothesis that weaknesses in the design and operation of the Vermont program underlie much of the problem of low reliability. A redesign of the program might incorporate, for example, more conventional (and narrower) definitions of types of tasks (such as "persuasive essays" and "narrative writing") and rubrics that are simpler and perhaps genre-specific.

In a forthcoming report, we will elaborate on these issues and will place them in the context of information about the implementation and impact of the assessment program in its second year of statewide implementation.

References

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). ***The reliability of scores from the 1992 Vermont portfolio assessment program: Interim report.*** **Santa Monica, CA: RAND Institute on Education and Training.**

Koretz, D., Stecher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). ***Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience*** **(CSE Tech. Rep. No. 371). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.**