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

The Reliability of Scores From the 1992 Vermont Portfolio Assessment Program

CSE Technical Report 355

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

RAND Institute on Education and Training/CRESST

February 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 RELIABILITY OF SCORES FROM THE 1992 VERMONT PORTFOLIO ASSESSMENT PROGRAM

Daniel Koretz, Daniel McCaffrey, Stephen Klein,

Robert Bell, and Brian Stecher

RAND Institute on Education and Training/CRESST

Summary

Since 1988, Vermont has been developing an innovative performance assessment program that relies substantially on portfolios of student work. The assessment program was designed to serve diverse goals: to provide rich data on student performance; to encourage better teaching and the adoption of higher standards; to coexist with Vermont's strong tradition of local control and innovation; and to encourage greater equity of educational opportunity.

RAND, as a part of the Center for Research on Evaluation, Standards, and Student Testing (CRESST), has been evaluating the Vermont Assessment Program since 1990. This interim report presents RAND's basic findings about the reliability of scores from the first statewide implementation of the portfolio program, in the 1991-92 school year. More detail about reliability will be reported at a later date.

An earlier report from the RAND/CRESST evaluation, *The Vermont Assessment Program: Interim Report on Implementation and Impact, 1991-92 School Year* (Koretz, Stecher, & Deibert, 1992), discussed the implementation and perceived impact of the program. Many Vermont educators found the program in its first year to be burdensome, and many pointed to aspects of the program that in their opinion needed improvement. Yet despite these difficulties, support for the program was widespread. Many educators reported that the system was a powerful lever for instructional change; many also reported that it had changed their own evaluations of students and increased their enthusiasm for teaching. Indeed, in about half of the schools investigated, local staff had already expanded the portfolio program beyond the

two grades (fourth and eighth) targeted by the state, and principals in numerous other schools expect to follow suit in the near term.

This report reflects a second component of the evaluation, in which RAND is examining the quality of the information yielded by the portfolio assessment. In this component, we focus not on the program's impact as an educational intervention, but rather on its quality as an assessment tool.

The reliability of scores depends in part on how they are constructed from data on student performance. In the Vermont portfolio program, students' writing was scored on five scoring criteria, each of which was scored on a 4-point scale. Pieces in the mathematics portfolio were scored on seven scoring criteria, again on 4-point scales. Although scores were combined in various ways across pieces of work, they were never combined across scoring criteria. Thus, for example, each student's "best piece" in writing was given five separate criterion scores but no single total score. *Our estimates of reliability reflect those decisions about scoring and reporting*. In some instances, alternative use of the data—for example, combining scores across criteria—could increase reliability. Several such possible alternatives and their effects on reliability are noted in the body of this report.

Given the ways in which portfolio scores were created, their "rater reliability"—that is, the extent of agreement between raters about the quality of students' work—was on average low in both mathematics and writing. Reliability varied, depending on subject, grade level, and the particular scoring criterion, and in a few instances it could be characterized as moderate. The overall pattern was one of low reliability, however, and in no instance was the scoring highly reliable.

A conventional statistic used to indicate the reliability of scoring is the "reliability coefficient," a measure of the extent to which two raters rank students' work the same. It ranges from 0.00 (essentially, no agreement beyond chance) to 1.00 (perfect agreement). Reliability coefficients of .70 or higher are not unusual for standardized performance assessments in writing (that is, assessments such as the Vermont Uniform Test of Writing, in which all students write responses to the same prompts). Indeed, the reliability of Vermont's own Uniform Test was in this range: .75 in fourth grade and .67 in eighth grade.

In the Vermont portfolio assessment, however, average reliability coefficients ranged from .33 to .43 (see Table 1). These are averages across all five scoring criteria in writing and all seven scoring criteria in mathematics. Although it may be unrealistic to expect the reliability of portfolio scores to reach the levels obtained in standardized performance assessments, these reliability coefficients are low enough to limit seriously the uses of the 1992 assessment results. The following sections of this paper provide additional detail and clarify the meaning of these coefficients.

Despite the unreliability with which individual students' work was scored, statewide average scores are quite reliable because of the large numbers of students represented. The 1992 averages, however, represent only those districts and schools that participated in the assessment program. Statewide estimates of the proportion of students reaching each score point are biased by the unreliability of scoring and cannot be reported.

One positive finding about the quality of the data is that teachers' own evaluations of their students' writing appear unbiased. That is, on average, teachers did not rate their own students' portfolios more positively than did volunteer teacher-raters.

At this point, we can only hypothesize about the causes of the low reliability of portfolio scores. However, characteristics of the scoring systems, aspects of the operation of the program, and the nature and extent of training are all plausible contributors to unreliability, and steps can be taken in each of these areas in an effort to increase the reliability of scores. For example, it

Table 1

Average Reliability Coefficients, Portfolio Scoring

	Grade 4	Grade 8
Mathematics Best Pieces	.33	.33
Writing Best Piece	.35	.42
Writing Remainder	.34	.43

Note. Because the scales used in rating the portfolios were only ordinal, Spearman correlations are reported throughout this memorandum. However, the more conventional Pearson correlations were only trivially different.

may prove helpful to simplify the scoring systems and to make training more uniform and rigorous. In addition, changes in reporting strategies, such as reducing the results to fewer, simpler statistics, could offer some improvement.

Background: The Vermont Assessment Program

Until recently, Vermont had no regular statewide assessment program. By the late 1980s, however, pressure was building to provide regular information on student performance, and by 1988, the state Department of Education began movement toward establishment of a statewide assessment system.

The deliberations that led to the decision to build the present, portfolio-based system are difficult to summarize succinctly because they were lengthy and involved many diverse people, including the Commissioner of Education (Rick Mills), the Department's then-Director of Policy and Planning (Ross Brewer), the governor, members of the state board, local board members, teachers, and others. Several persistent themes, however, were stressed by Mills, Brewer, and others working to build the system.[1] Ideally, the new system would:

- Avoid the distortions of educational practice that conventional test-based accountability created in some other jurisdictions;
- Encourage good practice and be integrally related to the professional development of educators;
- Reflect Vermont's tradition of local autonomy, "encourage local inventiveness, [and] preserve local variations in curriculum and approach to teaching" (Mills & Brewer, 1988, pp. 3, 5);
- Provide "a high common standard of achievement for all students" (Mills & Brewer, 1988, p. 3); and
- Encourage greater equity in educational opportunity.

[1] This description is based in large part on the first author's participation in meetings and discussions with Department of Education staff and others involved in building the assessment program. No single source summarizes the development of the program, but many of the points noted here have been described elsewhere. See, for example, Vermont Department of Education (1990, 1991a, 1991b) and Mills and Brewer (1988).

The basic outline of the assessment program emerged quite quickly. Eventually, the assessment would span a broad range of subjects, but the state decided to begin with assessments in writing and mathematics in Grades 4 and 8. The assessment would have three components: year-long student portfolios, "best pieces" drawn from the portfolios, and state-sponsored "uniform tests." The uniform tests would be standardized but not necessarily multiple-choice. A pilot implementation in a limited number of schools was conducted in the 1990-91 school year, and 1991-92—the year reflected in this report—saw the first statewide implementation of the program.

The details of the program, however, have been worked out only gradually. In contrast to the many states that either buy off-the-shelf tests or contract to have new tests built on a short schedule, the Vermont program was seen from the outset as a long-term and decentralized development effort. For example, in 1988, Mills called for mixing state-of-the-art assessment techniques with "emerging" techniques and warned that the development of the new program would be "a very long effort" (Mills & Brewer, 1988). In the Vermont program, the common call for a "bottom-up" approach was real rather than merely rhetorical; committees of teachers were created to take primary responsibility for the development of the assessment program, scoring criteria, and so on.

Thus, in both subjects, the so-called "pilot" implementation in 1990-91 was less a true pilot of a developed program than an integral part of the development effort. Indeed, in mathematics, even the first full statewide implementation in the 1991-92 school year would more accurately be considered a combination of a developmental effort and a pilot test, rather than a first implementation of a fully developed program. Some of the details of the scoring of best pieces in the 1991-92 statewide implementation, for example, were not resolved until spring of 1992, and ratings of entire mathematics portfolios have not yet been attempted on a large scale.

Reliability of the Writing Portfolio Assessment

Two sets of scores were obtained for each writing portfolio. One set was for the best piece; the other was for the remainder of the portfolio (called just the "remainder" below). Each set contained five scores, one for each of the five scoring criteria. Each criterion was scored on a 4-point scale. Scores for the

five criteria were not combined into a total score. The rating for the remainder was the rater's overall sense of the quality of the remainder as a whole in terms of the five scoring criteria; raters were not given instructions about methods for combining ratings across the pieces comprising the remainder. Classroom teachers provided the initial ratings of their students' portfolios, and volunteer teachers provided second ratings for a sample of students.

Rater reliability for both the best piece and the remainder varied but was generally low. Reliability was slightly higher for eighth-grade portfolios than for fourth-grade portfolios (Table 2). The lowest reliability was for the criterion of "voice" (.28 for the Grade-4 remainder), while the highest was for usage (.57 for the Grade-8 remainder).[2] In most cases, the rater reliability of the best piece was very similar to that of the remainder.

These low reliabilities stand in contrast to the Uniform Test of Writing, which was scored with fairly high rater reliability—.75 in Grade 4 and .67 in Grade 8. For a variety of reasons, such as the variability of tasks used, it may be unrealistic to expect a portfolio program to reach as high a level of reliability as a standardized performance assessment program. Extant research is not yet sufficient to indicate a reasonable target for reliability in portfolio programs. However, the reliabilities obtained in Vermont in 1992 are sufficiently low to limit severely the uses to which the results can be put.

Table 2

Reliability Coefficients, Writing Remainder

Scoring Criterion	Grade 4	Grade 8
Purpose	.33	.39
Organization	.31	.43
Details	.33	.41
Voice	.28	.37
Usage	.43	.57
Average	.34	.43

[2] Note that small differences in estimated reliability (between grades or among criteria) are unimportant and might reflect only chance.

A second way to present reliabilities, clearer to non-technical audiences, is the percentage of cases in which raters agree on the score given to a piece of work (or a portfolio). On average, raters agreed on the scores assigned to writing best pieces a bit less than half of the time (see Table 3).[3] The percentage agreement varied relatively little; in all but one case, the percentage agreement fell between 44% and 50%. Most differences were 1 point out of 4, but in some cases, raters disagreed by 2 or even 3 points.

In many cases, combining information will increase reliability. In the case of the writing portfolios, reliability can be increased somewhat by combining scores across the five criteria to create a single total score for the best piece or the remainder. The average reliability of the scores on each criterion for the fourth-grade remainder, for example, is .34 (Table 2). Replacing those five scores with a single total (or average) across the five dimensions would increase reliability to .45. In the case of the eighth-grade remainder, reliability would increase from .43 to .58. The effects on the rating reliability of the best pieces would be similar. Combining further by summing the best piece and remainder scores, however, would have only very small effects on reliability. Totaling across the five scoring criteria would similarly boost the rater reliability of the Uniform Test from .75 to .87 in fourth grade and from .67 to .82 in eighth grade.

Table 3

Writing Best Piece: Percentage of Students for Whom Raters Assigned the Same Score

Scoring Criterion	% Grade 4	% Grade 8
Purpose	45	45
Organization	44	50
Details	45	49
Voice	39	47
Usage	47	50
Average	44	48

[3] Some people have argued that we should classify ratings that are within one point of each other as "agreement." We did not do this because on a 4-point scale, even random ratings will produce a large percentage of cases in which raters are within one point of each other.

On the positive side, there was no sizable systematic bias in teachers' ratings of their own students. That is, on average, classroom teachers did not assign their students scores that were systematically too high or too low. On average, ratings by classroom teachers were virtually the same as those by volunteer raters. This is illustrated in Table 4, which provides the average ratings (on the 4-point scale) of fourth-grade best pieces by students' own teachers and second raters.

Reliability of the Mathematics Portfolio Assessment

The mathematics portfolio assessment operated differently than the writing assessment, and our reliability analysis differed accordingly. Students were instructed to cull five to seven best pieces from their portfolios. Each of these best pieces was rated on each of seven scoring criteria. As in writing, each criterion was scored on a 4-point scale. The ratings of the highest five pieces on each scoring criterion were then combined (by means of an algorithm designed by the mathematics committee) to produce a single composite score on each of the seven criteria. Thus, each student obtained seven scores—a composite score on each of the seven scoring criteria. These seven criterion scores were not combined into one or more total scores. As a result, we analyzed a single score per student per criterion in mathematics (unless otherwise noted, the composite score), in contrast to the two (best piece and remainder) analyzed in writing. In addition, in mathematics, classroom

Table 4

Average Ratings in Writing, Grade 4 Best Piece, by Classroom Teachers and Second Raters

Scoring Criterion	Classroom Teachers	Second Raters
Purpose	3.0	2.9
Organization	2.9	2.9
Details	2.8	2.8
Voice	2.7	2.7
Usage	2.8	2.9
Average Across 5 Criteria	2.8	2.8

teachers did not rate the work of their own students. Thus, in contrast to writing, we had no opportunity to appraise possible bias in teachers' ratings of their own students.

Rater reliability in the mathematics portfolio program was also generally low, about the same on average as the fourth-grade writing portfolios. Reliability coefficients average .33 in both grades (see Table 5). Again, reliability coefficients varied considerably depending on the grade and criterion, but they were not high in any instance. In three of the 14 instances (fourth-grade Understanding of Task, Outcomes, and Language of Mathematics), reliability coefficients fell below .30, which must be considered extremely low.

Rater reliability would have been improved modestly if students' overall scores had been a simple average of their scores on each criterion, rather than the composite formed by the mathematics committee's algorithm. In fourth grade, the average rater reliability (across all criteria) would have been .44 rather than .33 if students' overall scores had been a simple average of their scores on all five pieces (Table 6).

As in the case of the writing portfolios, combining information across the mathematics criteria would increase rater reliability. As noted, if the composite scores were replaced by simple averages across pieces, average rater reliability for the fourth-grade portfolios would be .44 (Table 6). If one

Table 5

Reliability Coefficients, Mathematics Composite Scores

Scoring Criterion	Grade 4	Grade 8
Language of Math	.23	.28
Math Representations	.33	.31
Presentation	.45	.42
Understanding of Task	.26	.35
How: Procedures	.44	.30
Why: Decisions	.40	.31
What: Outcomes	.23	.35
Average	.33	.33

Table 6

Reliability Coefficients, Mathematics Composite Scores and Simple Averages Across Criteria, Fourth Grade

Scoring Criterion	Composite	Average
Language of Math	.23	.34
Math Representations	.33	.39
Presentation	.45	.53
Understanding of Task	.26	.44
How: Procedures	.44	.52
Why: Decisions	.40	.48
What: Outcomes	.23	.39
Average	.33	.44

further combined, not just across pieces, but also across scoring criteria—so that each student's seven scores were replaced with a single total or average—rater reliability would be boosted to .57. A similar but somewhat smaller improvement would occur in eighth grade.

In both fourth and eighth grades, mathematics raters assigned the same score to a student's portfolio about 60% of the time (Table 7).[4] Mathematics portfolios showed higher agreement between raters than writing portfolios—despite similar or lower reliability coefficients—because mathematics scores were often highly concentrated at one or two points on the 4-point scales. This concentration of scores tends to depress reliability coefficients, even when raters agree much of the time.

A particularly extreme example of this concentration of scores appeared in eighth-grade results on the criterion "What: Outcomes of Activities." Scores on this criterion showed a reliability coefficient of only .35. Yet fully 89% of students received the same composite scores from two raters (Tables 7 and 8). The reason is that nearly all students were given a score of 1, the lowest possible score (see Table 8). Specifically, 92% of students (233 of the 253 for

[4] Because a number of different patterns of piece-level scores could produce any given composite score, the rate of agreement on the composite scores need not indicate the rate of agreement at the level of individual pieces.

Table 7

Mathematics Composite Scores: Percentage of Students for Whom Raters Assigned the Same Score

Scoring Criterion	% Grade 4	% Grade 8
Language of Math	52	49
Math Representations	55	56
Presentation	54	51
Understanding of Task	75	76
How: Procedures	66	62
Why: Decisions	47	49
What: Outcomes	81	89
Average	61	62

Table 8

Concentration of Scores in Mathematics: Grade 8, "What: Outcomes of Activities"

Scoring	A Percent Receiving Score From Rater 1	B Percent [of A] Receiving Same Score From Rater 2[a]
1	92	93
2	7	47
3	1	33
4	0	0
All Scores	100	89

Note. Reliability Coefficient = .35.

[a] This column gives the percentage of students receiving a given score by rater 1 who receive the same score from rater 2. Thus, 92% of all students received a rating of 1 from rater 1; 93% of the 92% given a score of 1 by rater 1 also received a score of 1 from rater 2.

whom two portfolio scores were available) were given a composite score of 1 by the first rater. Of those 233 students, 93% were given a score of 1 by the second rater as well. Although no other criteria showed this extreme a concentration

of scores, substantial concentration was shown by several. For example, in Grade 8, about 80% of students were given a rating of 3 on the criterion "Understanding of Task."

Although concentration of scores can produce a high rate of agreement even when the reliability coefficient is low, this agreement may reflect either reliable scoring or chance.[5] Given the low rater reliability shown on other criteria in the 1992 assessment data, it would not be reasonable to assume reliability in the case of criteria showing highly concentrated scores.

Reliability of Statewide Estimates of Performance

Despite the unreliability with which the work of individual students was scored, some statewide results from the 1992 portfolio assessment are reliable enough to report. The reason is that when results are aggregated over large groups of students, the error in assigning scores for individual students becomes relatively less important. However, the unreliability of scores in 1992 was so large that it limits the reporting of even statewide results.

Estimating the error in aggregate results from the Vermont assessment is complex, and several factors need to be taken into account:

1. *Sampling error*. In some cases (e.g., eighth-grade mathematics), we have scores for only a sample of the state's students. Moreover, each year's students are in a sense a sample from a larger pool of students flowing through the schools over time. This causes some uncertainty, for example, in estimates for schools or districts.[6]

2. *Clustering.* The portfolios of students within schools or classrooms are more similar than those of students from different schools. For example, in one sample of three schools, a total of 57 eighth-grade mathematics portfolios were scored, containing about 80 different tasks; only two of those tasks were common between two of the schools. Estimates of sampling error should be increased to take this into account.

---

[5] Because of the concentration of scores, the high rate of agreement in Table 8 is not that different from what would be obtained by chance. Given this concentration, if scores were paired at random, one would get an agreement rate of 85% ($.92^2 + .07^2 + .01^2$), as compared to the 89% agreement shown in Table 8. Similarly, the conditional probability that a student will receive a score of 1, given that another rater has already assigned a score of 1, is .93—only trivially different than the unconditional (overall) probability of 92%.

[6] The average score for a school is affected by year-to-year changes in the performance of successive cohorts of students, independent of any effects of schooling. Research has shown that the differences between "good crops" and "bad crops" of students can be sizable.

3. *Measurement error.* A key aspect of measurement error is the generally low rater reliability noted above; this decreases the reliability of statewide estimates.

4. *Biased estimates of proportions*. As explained below, the unreliability of scoring will generally result in too many students obtaining extreme scores. This has several effects on the uncertainty of statewide estimates.

The number of these factors that can be dealt with satisfactorily varies depending on the statistics used to report statewide results.

Average Scores

Although the Vermont State Department of Education has intended to report the proportion of students receiving each score rather than average scores, the unreliability of scoring is much more easily dealt with in the case of averages. The first three factors noted above (sampling, clustering, and measurement error) are taken into account in our estimates of the error bands for averages, and the fourth factor is not relevant.

The statewide average scores are reasonably precise, as illustrated by fourth-grade mathematics composite scores (Table 9). The first column provides the average score on each criterion, and the second column provides the width of a confidence band in each direction.[7] For example, the average score on "Language of Mathematics" was 1.7 out of a possible 4, and the margin of error extends .05 in either direction—that is, from 1.65 to 1.75. We found that these margins of error were in some instances as much as twice as large as they would have been with perfectly reliable scoring, but they were acceptably small nonetheless. Average writing scores showed trivially larger margins of error, but again they were small enough to be of little consequence (Table 10).

Despite the small margins of error, however, statewide averages for 1991-92 still have serious limitations. The most important is that some districts and schools—for example, the Burlington district—opted out of program. Accordingly, statewide averages must be interpreted as representing only participating schools and districts. Probably less important, scoring, particularly in mathematics, was subjected to various types of sampling—

[7] These confidence bands are twice the standard error, estimated by a school-level jackknife procedure to reflect clustering.

Table 9

Average Fourth-Grade Mathematics Composite Scores and Margins of Error

Scoring Criterion	Average	+/- Error
Language of Math	1.7	.05
Math Representations	2.3	.05
Presentation	2.5	.06
Understanding of Task	2.8	.04
How: Procedures	2.7	.05
Why: Decisions	2.5	.06
What: Outcomes	1.2	.04

Table 10

Average Eighth-Grade Writing "Remainder" Scores and Margins of Error

Scoring Criterion	Average	+/- Error
Purpose	3.0	.08
Organization	2.8	.07
Details	2.7	.07
Voice	2.8	.08
Usage	2.7	.06

some planned, some *ad hoc*—to compensate for the shortage of raters. We have not estimated the likely effects of the non-representativeness that might have resulted from factors such as these.

Proportions of Students at Each Score Point

Unfortunately, the low rater reliability discussed above greatly complicates estimating the proportion of students attaining each score point, despite the benefits of aggregating across a large number of students. Some of the proportions are likely to be systematically biased, and we cannot adequately estimate the margin of error around them.

Bias. When ratings are unreliable, scores tend to spread out more than they would with reliable scoring. That is, the distribution of observed scores, replete with measurement error, is more spread out than the distribution of underlying "true" scores. Too many students obtain extreme scores, and too few obtain scores near the middle.

In the case of the Vermont portfolio system, the resulting bias is substantial enough to undermine reporting of statewide proportions of students reaching each score point. We estimated "true" percentages for two criteria in fourth-grade mathematics, "Understanding" and "Presentation." In the case of Understanding, we estimated that the true proportion of students scoring either 1 or 4 was essentially zero, as opposed to the 2% and 1% observed (Table 11). The more substantial bias, however, was in the scores of 2 and 3. We estimated that the true proportion of students obtaining a score of 2 was about 8%, roughly half the 17% observed in the data. In the case of Presentation, the estimated true proportion at a score of 1 was 3%, rather than the 9% observed, and the true proportion at a score of 4 was zero rather than 4%. Estimates of true percentages, however, rest on assumptions that are somewhat risky, particularly when measurement error is as large as it was this past year in the Vermont program. Accordingly, we recommend against reporting either observed or estimated true proportions for 1992.

Margins of error. In some respects the systematic bias noted above makes a margin of error irrelevant for 1992 data, because we lack a reasonable estimate of each proportion to bracket with a margin of error. However, some

Table 11

Observed and Estimated True Proportions of Students at Each Score Point, Fourth-Grade Math, Understanding and Presentation

	Understanding		Presentation	
Score	Observed	True	Observed	True
1	2	0	9	3
2	17	8	39	45
3	81	92	47	51
4	1	0	4	0

discussion of the margins of error for proportions follows, in order to facilitate discussion of future reporting if scoring becomes more reliable.

In general, reporting of proportions will be more difficult than reporting of averages. In some cases, the margins of error will be relatively larger; moreover, accurate estimates of those error bands will be difficult to obtain. Two of the four factors noted above—sampling error and clustering—can be addressed adequately in the case of proportions. The other two, however—biased estimates of proportions and measurement error—are difficult to address adequately. Conventional estimates of the margins of error will be too small, because they do not take measurement error into account, and there is no simple and conventional method for correcting this. In addition, the bias in estimated proportions noted above will also bias estimates of error bands, but in inconsistent ways. Margins of error will be underestimated for observed proportions that are too extreme (too far from .50, because of the bias noted above) and will be overestimated for observed proportions that are not as extreme as they should be.

For illustrative purposes only, we have estimated the margins of error that would obtain for statewide proportions if scoring were reliable. Because of the problems noted above, these should not be considered the actual margins of error for the 1992 data. Rather, they illustrate what the margins of error are likely to be under more ideal circumstances.

The margins of error for statewide proportions are considerably larger than those for means (Table 12).[8] Thus, state-level reporting of proportions will necessarily be more imprecise than reporting of means. In the case of Presentation, for example, it would be more reasonable to present the proportion of students scoring 3 as "43% to 51%" or "roughly half," rather than "47%," because the margin of error is +/- 4 percentage points.

Error bands will be more problematic for reporting proportions below the state level, for example, at the level of schools or districts, because of the smaller numbers of students. To illustrate this, the following table provides the same margins of error for a proportion of 20%, assuming perfect rater

[8] These estimates take clustering into account by means of a school-level jackknife procedure. They do not take measurement error into account, and they do not adjust for bias in the proportions.

Table 12

Observed Proportions of Students at Each Score Point and Incomplete Margins of Error, Fourth-Grade Mathematics, Understanding and Presentation

	Understanding		Presentation	
Score	Proportion	+/- Error	Observed	+/- Error
1	2	1	9	2
2	17	3	39	3
3	81	3	47	4
4	1	0.5	4	1

reliability, for groups of different sizes (Table 13). Thus, if a school includes 30 eighth-grade students whose portfolios are scored, and 20% of them receive a certain score, the margin of error around the estimate of 20% extends from 13% to 33%.[9]

Table 13

Margins of Error for Proportions of 20%, by Number of Students, Perfect Reliability

Number of Students	+/- Error (Percentage Points)
15	10
30	7
45	6
60	5
100	4

[9] Because we assume no clustering below the level of schools—in part because of the small size of many Vermont schools—these are simple random sampling estimates of twice the standard error of a proportion.

Implications

The Vermont portfolio program faces substantial hurdles because of the unreliability of scoring documented here. Rater reliability is low enough to undermine the utility of 1992 scores for comparing groups of students (schools, districts, or other groups). Even when scores are aggregated enough to produce estimates with small measurement and sampling error—for example, statewide reporting of average scores—low reliability threatens their usefulness for gauging trends in performance over time, because it remains uncertain how an increase in the reliability of ratings will affect the distribution of scores even if true performance remains constant.

Low rater reliability could also hinder the instructional effects of the assessment program. Given that raters are teachers, low reliability suggests that teachers remain uncertain what skills are sought or, at the least, what performances constitute evidence that their students have mastered those skills. If teachers are inconsistent in their interpretation of desired performances, students will in turn receive inconsistent feedback on their work. Moreover, if the inconsistency among teachers indicates that some misinterpret the goals of the program, the feedback their students receive may be undesirable or incorrect.

Finally, the low reliability of scoring precludes many of the analyses that would otherwise be carried out to validate the assessment results. The portfolio assessment was conceived as a part of a larger assessment program, the pieces of which would be designed to measure different things. For example, in early planning discussions, some participants suggested that a uniform assessment is better suited than portfolios to measuring content coverage. Refinement of the program requires information about what things are actually being measured by each component of the system. The low rater reliability, however, severely constrains analysis of the information yielded by the portfolios.

The low reliability of scoring could have a variety of causes, and it is not clear at this time which factors are most important. However, three broad categories of factors are likely to contribute to some degree, and all warrant attention.

First, several aspects of the scoring systems may be contributing to the lack of reliability. Unclear or inconsistent terminology in the scoring rubrics could contribute error in scoring. For example, the writing system uses terms about frequency of occurrence to define generic points on the 4-point scale, but many of the scale points for specific criteria pertain to quality or extensiveness, not frequency of occurrence. Such inconsistencies may be interpreted differently by different raters and thus may contribute to unreliability. The complexity of the rubrics may also contribute. In the case of mathematics, raters are being asked to keep in mind 28 scale points (four on each of seven scales). This may be too complex an array for some (or all) raters to bear in mind. Raters may also be unable to distinguish five dimensions of performance (in writing) or seven (in mathematics). In the case of writing, the lack of a standard method for combining ratings across pieces to get a remainder composite score may have lessened reliability appreciably.

Insufficient training may also contribute to low reliability. The Vermont program has multiple goals, including both professional development and the production of reliable and valid information about achievement. Particularly at early stages in the program's development, these various goals sometimes conflict. For purposes of professional development, it is desirable to spread training (and responsibility) for scoring among as many teachers as possible. To increase the reliability of scoring, however, it may be necessary to concentrate resources available for training on a relatively small number of teachers in any given year, so that most of the trained teachers can reach a high level of proficiency in scoring. It is also possible that the specific nature of the training offered to Vermont teachers needs to be changed. For example, we have been told that many writing teachers were instructed that coming within one scale point was "close enough"; it is not, if one wants reliable scoring on a 4-point scale. However, we have not analyzed the nature of Vermont's training in any detail.

Finally, the nature of the portfolio assessment program—in particular, the lack of standardization of tasks and administrative conditions—may be undermining reliability. Many performance assessment programs that have obtained high levels of rater reliability are *standardized* assessments, in that students perform the same tasks under the same conditions. In such assessments, raters can be trained with benchmark sample performances of

the precise tasks used in the assessment. This cannot be done in the case of unstandardized performance assessments such as portfolio assessments. In the Vermont portfolio program, raters must be trained more generically and must stretch the rating framework to fit tasks that differ greatly on many dimensions. This may undermine reliability because raters may disagree about how the criteria should be stretched when applying them to new tasks that are dissimilar to those used in training. Moreover, the tasks are not administered under standard conditions, which may introduce inconsistencies in the interpretation of tasks and in judgments of performance. The higher rater reliability in the Uniform Test of Writing may indicate that variation among tasks is contributing to rater unreliability.

Lacking more information about the relative importance of factors such as these, we can offer only general suggestions about ways of improving reliability. Steps could be taken in four areas: the scoring systems, training, operation of the portfolio system, and reporting.

Scoring systems. The results of the Uniform Test suggest that at least in the case of writing, problems with the scoring systems may prove to have less impact than training and variations in tasks and administrative conditions. Nonetheless, refinements in the scoring systems may have an appreciable payback, and they will generally be easier and cheaper to effect than changes in other aspects of the system. We recommend that the rubrics be investigated for inconsistencies and lack of clarity and that raters be questioned about difficulties in applying them. The creation of composite scores in writing should be made systematic rather than impressionistic, and alternative methods of creating composites should be explored. We also recommend that the state Department of Education remain open to the possibility of simplifying the system to include fewer criteria.

Training. Despite the professional development goals of the program, we recommend that training be concentrated enough to bring raters to a higher level of performance. If larger groups are asked to rate portfolios for other reasons (for example, for professional development), their ratings should not be used for external reporting until they can be sufficiently trained. We also recommend that training be made uniform and that procedures be put in place to assess the competence of raters after training (and before statewide ratings).

Operation of the portfolio system. Over the long run, obtaining satisfactory levels of reliability may require substantive changes in the operation of the program—for example, more stringent restrictions on the types of performances allowed in the portfolios or on the conditions under which those performances are generated. Only additional experience and data will clarify in detail what changes should be considered, but some initial steps seem warranted now. It appears that further efforts are needed to clarify what genres of work are sought (or not sought) in mathematics. The further development of banks of exemplar tasks might be an important aspect of this effort.

Reporting of results. Changes could be made in the reporting of results to lessen unreliability or to mitigate its effects. Because averages (across students) will have smaller and more accurately estimated margins of error than the proportions of students reaching each score, the state Department of Education may wish to consider relying substantially on averages for comparative reporting and using proportions for supplementary detail. In addition, as noted earlier, replacing the current mathematics composite score with the simple average across pieces would increase reliability somewhat. Different approaches to pooling across pieces have different substantive advantages and disadvantages, but the effects on reliability may be a consideration worth weighing in deciding among them. Finally, as shown earlier, combining information can increase reliability, albeit at the cost of decreasing detail. The example used above—replacing criterion-specific scores with a single total score for each portfolio or best piece—is the most extreme case. Vermont educators may decide that such an approach loses too much detail, but there are intermediate levels of detail as well. For example, mathematics scores could be combined to provide two total scores per portfolio, one for communication and the second for problem-solving. Here again, substantive preferences should be weighed against the effects on reliability.

References

Koretz, D., Stecher, B., & Deibert, E. (1992). *The Vermont Assessment Program: Interim report on implementation and impact, 1991-92 school* year (CSE Tech. Rep. 350). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Mills, R. P., & Brewer, W. R. (1988). *Working together to show results: An approach to school accountability in Vermont.* Montpelier: Vermont Department of Education, October 18/ November 10.

Vermont Department of Education. (1990, September). *Vermont Writing Assessment: The pilot year.* Montpelier, VT: Author.

Vermont Department of Education. (1991a). *Looking beyond "The Answer": The report of Vermont's Mathematics Portfolio Assessment Program.* Montpelier, VT: Author [undated].

Vermont Department of Education. (1991b). *"This is my Best": Vermont's Writing Assessment Program, pilot year 1990-91.* Montpelier, VT: author [undated].