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

Can Portfolios Assess Student Performance and Influence Instruction? The 1991-92 Vermont Experience

CSE Technical Report 371

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

RAND Institute on Education and Training

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

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.

PREFACE

This report is intended for practitioners, researchers, and policymakers concerned with current performance assessment efforts and their effects on instructional quality. It describes the results of an evaluation of Vermont's statewide assessment initiative, a program that has garnered widespread attention nationwide because of its reliance on portfolios of student work. This report provides information about the implementation of the program, the effects of the reform on educational practice, the analytic challenges presented by the portfolio scoring process, the reliability and validity of portfolio scores, and the tensions between assessment and instructional reform. The Vermont experience has important implications for reforms that are underway or under consideration in other jurisdictions.

This project was conducted by RAND under the auspices of the Center for Research on Evaluation, Standards, and Student Testing (CRESST). The report presents the results of RAND's evaluation of the first statewide implementation of the Vermont portfolio assessment program, which was undertaken in the 1991-92 school year. This report integrates and adds to material presented in earlier reports and conference papers issued by the RAND study team during 1991-92.

The first four authors contributed to the preparation of all sections of this report. However, Brian Stecher was the principal author of Chapter 2, and Stephen Klein took primary responsibility for preparing Chapters 3 and 4. Dan Koretz was the principal author of the remaining chapters. Dan McCaffrey performed statistical analyses that are reflected throughout the report. Ed Deibert was responsible for many of the operational aspects of the study and also contributed material reflected in Chapter 2.

CONTENTS

FIGURES

TABLES

SUMMARY

Since 1988, Vermont has been developing an innovative assessment program in which student portfolios play a central role. Mathematics and writing portfolios were piloted in selected and volunteer schools in 1990-91, and the program was first implemented statewide in the 1991-1992 school year. RAND has consulted with Vermont about the development of the program since 1988 and has been evaluating the program since 1990 as part of the Center for Research on Evaluation, Standards, and Student Testing (CRESST). RAND's evaluation has encompassed the implementation of the program, its effects on educational practice, and the quality of the achievement data produced by the program.

This report describes the result of RAND's evaluation of the first statewide implementation of the program in 1991-92. This report integrates and adds to material presented in earlier interim reports (Koretz, McCaffrey, Klein, Bell, & Stecher, 1992; Koretz, Stecher, & Deibert, 1992) and in conference papers.

Background on the Vermont Assessment Program

Until recently, Vermont had no statewide assessment program. By the late 1980s, however, pressure began building to create one. Richard Mills, the Commissioner of Education, and W. Ross Brewer, then Director of Policy and Planning for the Vermont State Department of Education, initiated development of the current system in 1988. The program, like many other performance assessment initiatives nationwide, was intended to serve two purposes: (a) to provide useful information about student performance, and (b) to encourage improvement in teaching. Mills' and Brewer's goal was to build a system that would be consistent with Vermont's tradition of highly decentralized educational decision making while still encouraging "a high common standard of achievement for all students" (Mills & Brewer, 1988). They explicitly acknowledged that the system would necessarily go beyond current technologies and would require a long period of development.

The system that emerged has as its centerpiece portfolios of student work that are collected by classroom teachers but are scored using criteria that are

consistent across the state. The guidelines for the operation of the portfolio program and the criteria for scoring student work have been developed by committees consisting largely of volunteer teachers. Teachers are given only limited guidelines for the compilation of portfolios—for example, the number of pieces that should be included and the broad categories of work from which the pieces should be drawn. The portfolios are complemented by a system of "Uniform Tests" that are standardized in terms of content and administrative conditions but are not restricted to the multiple-choice format. Unlike the portfolios, the Uniform Tests have been developed largely by external contractors. Although the assessment system is intended to cover many subjects, at present it is limited to writing and mathematics in Grades 4 and 8.

In mathematics, the system (as implemented in 1991-92) required students and teachers to construct a portfolio for each student comprising five to seven "best pieces." These pieces were to be of three types: puzzles, applications, and investigations. A sample of portfolios from each school was submitted to one of seven regional meetings for scoring by other teachers. Each piece was scored on seven dimensions, four pertaining to problem solving and three pertaining to communication. All dimensions were scored on 4-point analytic scales. Composite scores (across the best pieces) were calculated for each dimension, but the state did not report a total score (combined across dimensions). A subsample of portfolios was scored twice for analysis of reliability.

The writing portfolio system operated differently. Students were asked to include in their portfolios six to eight pieces (depending on their grade level) from several different categories of work. (These categories were broader than traditional genres; examples include "a poem, short story, or personal narration" and "a personal response to a book, event, current issue, mathematical problem, or scientific phenomenon.") The student had to designate one piece from any category as the "best piece." The best piece was scored on five dimensions; again, all five were 4-point analytic scales. Raters were asked to score the remaining pieces in each portfolio as a set on each dimension and were not required to score each piece separately in arriving at scores for the set. The initial rater was the student's own classroom teacher, and a sample of portfolios was submitted for rescoring by a second rater.

Implementation and Impact

In 1991-92, information about implementation and impact was gathered from teachers and principals. Teachers responded to a questionnaire that focused on the mathematics portfolios. Their responses indicated that participation in the mathematics portfolio assessment was extensive and that the program appears to have had a positive impact on instruction.

- **Virtually all fourth- and eighth-grade math teachers received state-sponsored training in the use of portfolios. They generally rated this training as effective.**
- **Nearly all fourth- and eighth-grade students compiled portfolios of their mathematics work.**
- **Teachers reported devoting substantially more attention to problem solving and communication in teaching mathematics as a result of the program.**
- **Teachers reported some changes in mathematics instructional practices; for example, students spent more time working in small groups and in pairs.**

The mathematics portfolio assessment program produced other benefits as well. Portfolios were an impetus for positive changes in teacher and student attitudes regarding mathematics and learning. Many teachers reported that both they and students were generally more enthusiastic about mathematics as a result of the portfolios. The portfolios also provided teachers with new perspectives on students' abilities.

However, there were also shortcomings in implementation, as might be expected with a reform of this scale and novelty. The most serious problem was continuing confusion on the part of many teachers about the purposes of the mathematics portfolios and the proper practices to use to implement the assessment system. Most teachers felt they were unprepared to use the portfolios on at least some occasions. Teachers also raised concerns about the lack of information from the state and the rapid speed of the reform.

In addition, the decentralized nature of the reform encouraged teachers to adopt idiosyncratic practices regarding mathematics portfolios. Teachers assigned tasks of varying novelty and complexity. They adopted different rules

regarding revisions and the amount of assistance pupils could receive when preparing portfolio pieces. These contextual differences probably affected student performance and, as a result, the validity of comparisons based on portfolio scores.

Principals echoed these concerns and raised other issues relevant to both the writing and mathematics portfolios. They noted that the positive changes came at a price. The writing and mathematics portfolios required a sizable commitment of preparation and classroom time from teachers. The program also posed sizable demands on school resources for release time and substitute teachers. Furthermore, the portfolio assessment generated some negative attitudes on the part of teachers and principals. Both groups perceived the time and resources demands to be burdensome. Teachers also were anxious about scoring and the potential uses of scores, particularly for comparative purposes.

Although Vermont educators varied markedly in their opinions of the assessment program, a widespread view was that it was a worthwhile burden. That is, the demands of the program were seen as large and burdensome, but support for the program as a tool of instructional reform was high nonetheless. Perhaps the most telling sign of support was the fact that in roughly half of the schools in which we conducted interviews, the use of portfolios had already been expanded in some form beyond the grades or subjects included in the state program.

Overall, Vermont appears to be building momentum toward greater acceptance and more efficient use of the portfolios by responding to needs with targeted training. However, there are still important gaps between the ideal and the reality. The most significant gaps relate to standardization of basic practices and the establishment of a common understanding of fundamental constructs, such as problem solving and mathematical communication.

Reliability of Portfolio Scores for Individuals

Although the Vermont program was not designed to produce student-level scores for external use (that is, for use outside of students' own schools), the scores assigned to students are the basic building block of the system. The quality of student-level scores influences the adequacy of aggregate scores and limits the uses to which the assessment results can be put.

An essential component of reliability is "rater reliability" or "rater agreement"—that is, the extent to which different raters agree about the score that should be assigned to a given piece of work. The rater reliability of portfolio scores in both mathematics and writing was very low. The Vermont system was designed to provide separate scores on each scoring dimension. For individual pieces on a single dimension, the correlations of scores between raters ranged from .28 to .57; the average correlations (across all dimensions) ranged from .34 to .43. (A correlation of 1.00 indicates perfect agreement, while 0.00 indicates a total lack of a systematic relationship between raters.) The percentage of cases in which raters agreed on a score was generally not much higher than expected by chance. Summing across all pieces within a portfolio on a single scoring dimension raised the average reliability coefficients only slightly, to a range of .38 to .49. Vermont did not combine scores across dimensions because to do so would have been inconsistent with the program's goal of providing dimension-level scores as instructional feedback. However, even summing across dimensions as well as pieces to get a single total score would have raised the average reliability coefficients only to a range of .49 to .60.

Quality of Aggregate Scores

The Vermont system is intended to provide various types of aggregate scores, including the proportion of students statewide reaching each score point on each scoring dimension and comparative data about districts or schools.

In 1991-92, many of the aggregate statistics the state wished to report were not of high enough quality to use, and the Vermont State Department of Education accordingly declined to release them. The low quality of aggregate scores stemmed in part from the low rater reliability of the scores assigned to individual students.

Because of the relatively large number of students involved, statewide average scores on each dimension were quite reliable despite the low reliability of scoring. For example, in the fourth grade, the average score (on a scale from 1 to 4) on the dimension "Math representations" was 2.3; the confidence band for this average was approximately 2.25 to 2.35. However, this confidence band takes into account only sampling error and the reliability of ratings. It does not consider other factors that might threaten the reliability of scores, such as

limited generalizability of performance across the small sample of tasks scored for each student.

On the other hand, school-level average scores were too unreliable to use. The unreliability of scoring contributed to this problem but was not its main cause. Rather, the unreliability of school averages stemmed mainly from the small size of many Vermont schools, particularly in the fourth grade.

Furthermore, the proportion of students reaching each level could not be reported even for the state as a whole. One effect of unreliable scoring is to spread out the distribution of scores, so that too many students receive high and low scores and too few receive scores near the middle of the distribution. The resulting bias in estimates of the proportion of students receiving each score was large but could not be estimated well enough to permit statistical correction.

Validity

Rater reliability is an essential but insufficient criterion for evaluating assessment results. The ultimate question is the validity of the results—that is, the extent to which they support the conclusions people base on them. Unreliable scoring undermines validity, but reliable scoring does not guarantee it.

The issue of validity was largely moot in 1991-92 because of the low rater reliability of scores, but we explored what conclusions about validity might have been warranted if reliability of scoring had been better. This was done by "disattenuating" relationships in the data to estimate what would be found if scoring were perfectly reliable—a conventional approach, but risky in the case of Vermont's data because the measurement error was very large. Hence our findings can be considered only exploratory.

In general, our analyses of validity evidence were unpersuasive. A few of the patterns one would expect did appear in the data; for example, disattenuated correlations between writing portfolio scores and scores on the Uniform Test of writing were consistent with other assessments. However, many of the expected relationships failed to appear. For example, mathematics portfolio scores would be expected to correlate more highly with scores on the mathematics Uniform Test than with scores on the writing Uniform Test. In the main, this was not the case. Perhaps more important, evaluation of validity was hampered by the lack

of a sufficiently clear definition of the attributes the portfolios are intended to measure. Without such a definition, one cannot adequately use other information about performance to gauge the validity of scores.

Our data about program implementation also raise concerns about validity. Teachers reported large variations in key aspects of implementation, such as policies pertaining to revision of students' products. These variations in implementation might mask differences in student's capabilities or generate spurious differences where none are warranted.

Implications

The Vermont assessment program has attracted nationwide attention. A critical question for observers from outside Vermont is the extent to which the findings described here have implications for their current or planned performance assessment programs.

The Vermont program is in some respects unusual. Vermont's program emphasizes portfolios, which are unstandardized performances, while many other current programs rely on standardized performances, such as uniform writing prompts or standardized open-ended mathematics questions. Moreover, the Vermont program, like any other, has idiosyncratic elements (such as the particular rubrics used in scoring).

Nonetheless, the Vermont program has much in common with many other current performance assessment programs, and the Vermont experience has important implications for them. It shares with many other programs the dual, fundamental goals of measuring student performance and sparking improvements in instruction, and the Vermont program's specific instructional objectives (such as an increased emphasis on problem solving in mathematics) are shared by programs nationwide. Moreover, leading proposals would create new programs that resemble the Vermont program. For example, the New Standards Project has long called for extensive reliance on unstandardized performances. The Vermont experience can provide guidance for the design of these programs and can help set expectations for their performance.

Expectations for Quality of Measurement

In 1991-92, the quality of the data about student performance yielded by the portfolio program was so low that it severely restricted the appropriate uses of scores. Although unique aspects of the Vermont program undoubtedly contributed to this problem, some of the likely causes will arise in other programs as well. For example, there are at least three factors that are likely contributors to the unreliability of scoring: inadequate scoring rubrics, insufficient training, and the lack of standardization of tasks. Unstandardized tasks are likely to complicate scoring substantially in most programs that rely on them, particularly in subjects such as mathematics and science in which tasks are likely to vary greatly. Training of teachers and raters is likely to be an arduous task in many large-scale systems, particularly those that aim to change instruction markedly. Similarly, the limited generalizability of performance across tasks that we found in the mathematics portfolios is consistent with findings from diverse performance assessments, both standardized and not, and threatens the validity of inferences based on small numbers of tasks. The wide variations in program implementation we found are likely to occur in other programs that attempt to integrate assessment and instruction and will undermine the validity of many intended uses of scores.

Expectations for Impact on Educational Practice

The Vermont experience provides grounds for optimism that performance assessment programs may be able to influence instruction substantially. At the same time, it offers reasons to temper the optimism that is currently so widespread. Even after years of development, a year of piloting, and a full year of statewide implementation, the program had made only partial strides toward its instructional goals. For example, it had not yet provided mathematics teachers with a sufficiently consistent interpretation of performance goals. Moreover, the partial success attained to date has come at a high price in time, stress, and money. It seems clear that the performance assessment itself is only part of what is needed; it must be coupled, for example, with extensive and continuing professional development.

Tensions Between the Goals of Assessment Programs

Many performance assessment programs and proposals share with the Vermont program the dual goals of improving instruction and measuring student performance. To some degree, those two goals conflict. For example, standardization of both tasks and administrative conditions will generally improve the quality of student performance data, but standardization may hinder the integration of assessment with instruction and lessen teachers' feelings of ownership and commitment. The improvement of these programs will often require not only that steps be taken to lessen the tensions between these goals, but also that educators make difficult compromises between them. That is, they may need to choose how much of a decrement in measurement quality they are willing to accept to facilitate instructional improvement, and vice versa.

This fundamental tension manifested itself concretely in the Vermont program. We noted above that the wide variations in the implementation of the program, such as variations in policies toward the revision of students' work, threaten the validity of comparisons based on portfolio scores. To some degree, however, these variations in practice may be instructionally beneficial. For example, many teachers would argue that as students become more able, they should become more autonomous and should receive less assistance and structure to guide their revisions. To eliminate such variations may be instructionally undesirable, but to leave them in place undermines the validity of comparisons based on portfolio scores. The tension between the program's two fundamental goals also manifested itself in designing the scoring system. For the sake of professional development, Vermont opted to train all teachers to score and to use all teachers who volunteered as raters. To increase reliability, however, one would want to limit scoring to the number of raters that can be trained to a high level of accuracy.

Requirements for Evaluation

Programs that hold people accountable for scores on assessments are not self-evaluating. That is, increases in scores on the assessments for which people are held accountable are not sufficient to indicate that the programs are meeting their goals. This fact is now widely recognized in the case of programs using traditional, multiple-choice tests, but it is no less true of performance assessment programs. These new programs are intended to have pervasive

effects throughout the educational system, and evaluating these varied effects requires collecting information about a wide range of questions.

Explicit investigation of program implementation—and variations in implementation—is important to provide corrective feedback and to explore matters of equity. Changes in instruction must be investigated directly and cannot be inferred adequately from performance on the assessments. Examination of reliability must go well beyond simple rates of agreement among raters. It will be important to explore both the causes and the effects of unreliable scoring, when it is found. If and when reliable scoring is attained, it will be essential to research score reliability—that is, the generalizability of performance across tasks (and other facets) as well as raters—and other evidence of validity. Finally, to validate the new assessments and to obtain solid estimates of changes in student learning, evaluators will need to administer additional measures of the domains that the new assessments purport to measure.

Conclusions

The experience of the Vermont portfolio program to date suggests the need for patience, moderate expectations, and ongoing, formative evaluation. Neither of the basic goals of the program—using complex performances to measure student performance and utilizing a performance assessment program to spur instructional improvement—can be met easily or quickly. Moreover, the Vermont experience illustrates the tensions between these goals and the need to make difficult trade-offs in compromising between them.

ACKNOWLEDGEMENTS

Many people contributed to the work presented in this report. We are grateful to Richard Mills, Commissioner of Education in Vermont, and W. Ross Brewer, formerly Director of Policy and Planning for the Vermont State Department of Education, for providing us the opportunity to conduct this research, for their unwavering commitment to evaluation of their program, and for their consistent support of our efforts. Sue Rigney, current Director of Educational Policy and Planning, assisted the effort in numerous ways. Literally hundreds of Vermont educators generously contributed by filling out questionnaires, participating in interviews, providing logs of assignments, and permitting observations of classes. Robert Bell and Ellen Harrison contributed substantially to the statistical analysis, and Eric Hamilton made important contributions to the analysis of portfolio contents and teacher questionnaires. Bill Thompson of TBA Consulting Group provided help throughout, including detailing the procedures used in scoring mathematics portfolios. Mary Ann Minardo assisted our effort in many ways, including interviewing principals and teachers and observing classes. Sarah Blumer, Tracey Cochran, Mariellen Pestana, and Tina Corasaniti also interviewed teachers and observed classrooms. Advanced Systems for Measurement in Education provided data files, and Mark Wiley at Advanced Systems helped in numerous ways. Nancy Rizor, Rose Marie Vigil, and Diana Millar prepared the manuscript.

CHAPTER 1. INTRODUCTION

Since 1988, Vermont has been developing an assessment program that is at the cutting edge of innovation in large-scale assessments. Although a rapidly growing number of statewide assessment programs incorporate some form of performance assessment, the Vermont program is unusual among them in that its centerpiece is student portfolios and "best pieces" drawn from them. Portfolio-based assessment is not new, but Vermont was the first state to make portfolios the backbone of a statewide assessment program. The Vermont program is also unusual in the degree to which it is "bottom up": many aspects of the assessments in each subject are worked out iteratively by committees of teachers, and classroom teachers retain wide latitude in implementing the program.

The Vermont State Department of Education selected 48 schools to pilot fourth- and eighth-grade assessments in writing and mathematics during the 1990-91 school year, and 90 other schools asked to participate to varying degrees in the pilot efforts that year. The first statewide implementation of the assessment in those two grades and subjects was conducted in the 1991-92 school year.

RAND has consulted with Vermont about the development and eventual evaluation of the assessment program since August, 1988. Since 1990, RAND, as part of the Center for Research on Evaluation, Standards, and Student Testing (CRESST), has been carrying out a multifaceted evaluation of the assessment program and its effects. This monograph reports the findings of the RAND/CRESST study through the 1991-92 school year.

Background on 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 (Richard 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 appeared to have created in some other states;**
- **Encourage good practice and be integrally related to the professional development of educators;**
- **Reflect the Vermont 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"; and**
- **Encourage greater equity in educational opportunity (Mills & Brewer, 1988, p. 3).**

Those responsible for the nascent program were aware of the difficulties inherent in having an assessment program serve many functions at once and had been warned that some of their goals for the program pointed to different assessment designs. For example, a system designed to provide rich information about students and positive incentives for teachers might look very different from a system that was designed primarily to provide highly comparable information across schools.[2] The system was intended to be a compromise among the many goals for the system; for example, it should provide reasonable comparability across schools, but not at the cost of stifling good practice and local innovation.

---

[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 State Department of Education, 1990, 1991a, 1991b, and Mills and Brewer, 1988.)

[2] Dan Koretz, presentation to Commissioner Mills, Governor Kunin, and others, August, 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 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 and Brewer 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). 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, the evaluation of which is reported here, would most accurately be categorized as a combination of a developmental effort and a pilot test, rather than as an initial implementation of a fully-planned 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 portfolios have not yet been attempted on a large scale.

Primary responsibility for the development of the portfolio and best-pieces components of the program was given to state-sponsored committees of teachers. These committees worked independently of each other, so the program evolved differently in writing and mathematics.

Mathematics

As implemented in 1991-92, the mathematics program required that students and teachers cull from each student's portfolios a set of five to seven "best pieces." Teachers were requested to include in this set of best pieces exemplars of three types of problems: puzzles, investigations, and applications. According to the *Teacher's Guide* (Vermont State Department of Education, 1991c), *puzzles* are tasks that "require students to identify and explore approaches to nonroutine problems... [where] most of the problem deals with logic and reasoning." For example, "With a seven minute hourglass and an

eleven minute hourglass, how could you time the boiling of an egg for 15 minutes?" *Applications*, by comparison, "require students to apply knowledge they already possess." For example, "A mature tree can utilize 13 lbs. of carbon dioxide in a year. The average car spews out 4000 lbs. of carbon dioxide in a year. How many mature trees would you need to utilize this much carbon dioxide?" (The problem continues, exploring the relationship of fuel efficiency and carbon dioxide emissions.) *Investigations* include "explorations, data collection and analysis or some level of research that leads to conclusions." For example, "Given a piece of graph paper, determine the size of a square to be cut out of each corner of the graph paper that will allow you to fold the graph paper into an open box with the greatest volume. Try squares from 1 to 9 units. Then determine the volume of each open box after you have folded the paper. Collect the data in a data table of your design. Graph the results. Write up your results." The *Resource Book* provided to each teacher (Vermont State Department of Education, 1991d) contained many samples of each type of task.[3]

The best-pieces sets of a sample of students from each participating classroom were sent to one of seven central sites for scoring by groups of volunteer teachers. (The samples from each class were selected by the state's assessment contractor, not the classroom teachers.) Training in scoring had been offered statewide before the scoring sessions, but the volunteer raters varied substantially in their level of training, and supplementary training was provided at the beginning of each three-day scoring session.

All of the best pieces were graded on 4-point scales on seven dimensions; four classified as problem solving and three as communication. The seven dimensions, and the scale points on each for individual pieces, were as follows:

- **Understanding the task: 1. totally misunderstood; 2. partially understood; 3. understood; 4. generalized, applied, extended.**
- **How – Approaches/procedures: 1. Inappropriate or unworkable; 2. appropriate some of the time; 3. workable; 4. efficient or sophisticated.**
- **Why – Decisions along the way: 1. no evidence; 2. possible; 3. inferred with certainty; 4. shown/explicated.**

---

[3] The *Resource Book*, which was prepared by a committee of teachers and distributed by the state Department of Education, contains a collection of activities appropriate for inclusion in mathematics portfolios.

- **What – Outcomes of activities: 1. no extension; 2. observations; 3. connections, applications; 4. synthesis, generalization, abstraction.**

- **Language: 1. No or inappropriate; 2. some of the time; 3. most of the time; 4. rich, precise, elegant.**

- **Representation: 1. No use; 2. use; 3. accurate and appropriate; 4. perceptive.**

- **Presentation: 1. not clear; 2. some clear parts; 3. mostly clear; 4. clear.**

The ratings on the individual pieces were then aggregated to provide an overall rating of the entire set of best pieces on each of the seven dimensions. However, the nature of the scales for individual pieces led the mathematics committee to reject a simple averaging across pieces. A set of three 4s ("perceptive") and three 2s ("use") on "Representation" do not average to a 3 ("accurate and appropriate"). Accordingly, the mathematics committee developed a "holistic" rating system in which the four scale points for the aggregate rating were defined differently than those for the individual ratings, and a fixed algorithm (based on the number of instances of different scores on the individual pieces) was used to create the aggregate scores.

This system for aggregating scores was too detailed to warrant a full description here, but a brief description of the ratings on the "Understanding the task" criterion illustrate the general approach. The four scale points for the holistic rating of the set of best pieces were:

1: Totally misunderstood in more than half of the best pieces.

2: May have understood or read beyond the surface problem in some instances but only partially understood more than half of the time or totally misunderstood the problem in two or more instances.

3: Understood or read beyond the surface problem most of the time but partially understood or misunderstood in some instances.

4: Understood the task most of the time and read beyond the surface problem at least a couple of times.

Each of these descriptions was accompanied by as many as 10 combinations of scores from individual pieces that would produce the aggregate rating. Only the five best pieces were counted, regardless of whether the student included one or two more (as guidelines permitted).

The mathematics portfolio assessment was accompanied in the 1991-92 statewide implementation by a sample-based administration of the state's new Uniform Test of mathematics. The Uniform Test was a matrix-sampled, mixed-format test, combining multiple-choice and open-ended items. Unlike the portfolio assessment, the Uniform Test was designed and scored by Advanced Systems for Measurement in Education, Vermont's testing contractor during the 1991-92 school year.

Writing

The design of the writing assessment, which was largely completed during the 1990-91 pilot year, is substantially different from that of the mathematics assessment. In writing, students' portfolios must include a set number of pieces of specified types, the entire portfolio is rated, and a single best piece is chosen.[4]

In Grade 4, each student's portfolio must include:

1. **A table of contents.**
2. **A single best piece, which is selected by the student, can come from any class and need not address an academic subject.**
3. **A letter explaining the composition and selection of the best piece.**
4. **A poem, short story, or personal narration.**
5. **A personal response to a book, event, current issue, mathematical problem, or scientific phenomenon.**
6. **A prose piece from any subject area other than English or language arts.**

The requirements for eighth grade are the same except that the portfolio must include three prose pieces.

The best piece, the rest of the portfolio, and performance on the Uniform Test of writing (which is a direct writing task using standardized conditions and a single prompt) were all scored on the same five dimensions:

- **Purpose;**
- **Organization;**
- **Details;**

---

[4] The following description is taken largely from *"This is my Best:" Vermont's Writing Assessment Program, Pilot Year 1990-91* (Vermont State Department of Education, 1991b).

- **Voice/Tone; and**
- **Usage/Mechanics/Grammar.**

A single 4-point scale was used with all five dimensions, labeled as: 1. rarely; 2. sometimes; 3. frequently; 4. extensively. The descriptions of the scale points, however, are generally phrased in terms of *quality* or *extensiveness*, not frequency. For example, in the case of purpose, the description of "sometimes" is:

- **"Attempts to establish a purpose.**
- **Demonstrates some awareness of audience and task.**
- **Exhibits rudimentary development of ideas" (Vermont State Department of Education, 1991b, p. 6).**

In the 1991-92 statewide implementation, teachers scored their own students' portfolios and best pieces. Advanced Systems scored the Uniform Tests and arranged for a sample of portfolios to be drawn from each class for "moderation"—that is, to be scored by an external panel of teachers so that the scores of participating teachers could be calibrated to a common standard.

The RAND/CRESST Studies

The characteristics of the Vermont assessment program required that the RAND/CRESST evaluation be broad in scope. The RAND/CRESST evaluation is a series of interrelated efforts designed to gather information about:

- **The implementation and operation of the program at the school and classroom level;**
- **The quality of measurement (including reliability and validity); and**
- **Effects on instruction and on other aspects of schooling.**

These questions have been addressed with a variety of methods, including questionnaires administered to teachers, interviews of teachers and principals, classroom observation, qualitative analysis of student portfolios, analysis of scoring methods and rubrics, questionnaires administered to scorers, and analysis of student-level and school-level scores.

Our initial efforts focused primarily on the mathematics portfolio program. The mathematics program represented more of a qualitative break with past

instructional practice than did the writing program. In addition, large-scale direct assessments of writing are quite common, even if large-scale portfolio assessments are not, and there is a research literature spanning decades about the characteristics of such assessments. In contrast, those building the Vermont mathematics assessment were largely plowing new ground, and their experiences could have important implications for the nationwide effort to develop direct assessments in mathematics. Accordingly, many of the results reported here pertain specifically to the mathematics program, while others pertain to both subjects.

The RAND/CRESST evaluation is formative. Our expectation, like that of the state Department of Education, is that the program will require a long period of development. Our evaluation is designed to monitor that process and to provide frequent corrective feedback along the way. The results reported here, which are limited to the pilot year (1990-91) and the first year of statewide implementation (1991-92) reflect only the initial stages of the program's evolution.

The Contents of This Report

Because of the state's need to use the results of the RAND/CRESST study for political decisionmaking and program design, the results of the 1990-91 and 1991-92 studies have been released piecemeal. This report incorporates material released earlier and adds information not previously published. One interim report, released in the summer of 1991 (Koretz, Stecher, & Deibert, 1992, often labeled "RAND I" by educators in Vermont), provided initial findings on implementation and impact from our questionnaires and interviews with principals. Chapter 2 of this report summarizes those findings and adds information from our interviews of teachers and our qualitative analysis of portfolios. A second interim report, released in December 1992 (Koretz, McCaffrey, Klein, Bell, & Stecher, 1992, often called "RAND II"), provided basic information on the reliability of scores. Chapters 3 and 4 of this report incorporate information from that report on the reliability of student scores and add additional detail, including a generalizability-theory analysis. Chapter 5 of this report discusses implications of the student-level rater reliability for the reporting of aggregate scores. Chapter 6 describes preliminary analyses of validity. Chapter 7 discusses other implications of the data reported here.

CHAPTER 2. IMPLEMENTATION AND IMPACT

Introduction

Our examination of the implementation and impact of the Vermont portfolio assessment program began when the portfolios were introduced on a limited pilot basis in 1990-91 and continued during the 1991-92 school year, when portfolios were used by all fourth- and eighth-grade mathematics and writing teachers in most of the state's districts. We focused our investigation primarily on the mathematics portfolio program because it represented a clearer break from extant practice and was, from a national perspective, more unusual. Unless otherwise noted,[1] the results reflect teachers' and principals' responses to the mathematics portfolios.

Portfolios are seen both as evaluative tools and as levers to reform mathematics curriculum and instruction. In evaluating the implementation of the program, we tried to determine how well this vision was instilled through training and instantiated in classroom practices.

The broad evaluation questions we sought to answer were:

- **How well was the portfolio assessment program implemented?**
- **What effects did the portfolio assessment program have on schools and classrooms (including effects on the content of the curriculum, the style and method of instruction, and the attitudes of teachers and students)?**
- **What burdens did the portfolio assessment place on schools and classrooms?**

The results reported below represent only a snapshot of an innovation in its early stages; subsequent evaluation of the program in later years will reveal more about its long-term effects.

---

[1] **Primarily in the discussions of implementation problems, teacher attitudes and the burdens imposed by the portfolios.**

Implementation

Data from the teacher questionnaires and the principal and teacher interviews permit us to examine three broad questions related to implementation:

- **How well-prepared were teachers to use the mathematics portfolios?**
- **In what ways did teachers use the mathematics portfolios, and were portfolio practices similar across teachers?**
- **What were the major implementation problems encountered by teachers and principals in implementing the portfolio program in both mathematics and writing?**

Teacher Preparation

Vermont provided teachers with many opportunities to learn about the mathematics portfolio assessment program and prepare for its implementation. During the pilot year (1990-91), the state sponsored a statewide orientation meeting in the fall and a series of regional workshops during the school year. In 1991-92, inservice training was expanded to include summer and fall mathematics institutes and preparation-for-scoring workshops during the year. All teachers were given a *Resource Book* with sample tasks and a *Teacher's Guide* with operational guidelines.[2] In addition, regional networks were established with consultants to provide supplemental training and support at the grassroots level.

Survey and interview responses presented a somewhat inconsistent picture of teachers' preparation to use the mathematics portfolios. Teachers indicated they were satisfied with the preparation they received, but they also said they continued to encounter problems they were not prepared to address. Specifically,

- ***Teachers generally said the state-sponsored mathematics portfolio training prepared them adequately.*** **Virtually all respondent mathematics teachers attended one or more state-sponsored training sessions, and approximately three-quarters of the teachers**

---

[2] **Vermont State Department of Education, *Vermont Mathematics Portfolio Project Resource Book*, Montpelier, September, 1991; and Vermont State Department of Education, *Vermont Mathematics Portfolio Project Teacher's Guide*, Montpelier, September, 1991.**

said these sessions prepared them adequately to work with the portfolios. Similarly, three-quarters of those teachers who received supplemental network-level training thought it prepared them at least adequately to work with the portfolios. Most teachers used the *Resource Book* distributed by the state as a key source of mathematics portfolio tasks during the year.[3]

- ***Nevertheless, teachers felt under-prepared to implement the mathematics portfolios at least some of the time.*** Approximately three-quarters of the teachers said they occasionally or frequently felt they did not have enough training to use the mathematics portfolios. Teachers commented that training left too many unanswered questions, was not extensive enough, and did not address the problems of scoring. A few principals also felt that the training was inadequate; they noted particularly the lack of training in using the portfolio scoring criteria.

Mathematics Portfolio Use

The Vermont portfolio assessment was designed to "encourage local inventiveness, [and] preserve local variations in curriculum and approach to teaching" (Mills & Brewer, 1988, pp. 3, 5). The official guidelines for the portfolios deliberately leave much to the determination of local teachers. As a result, it is important to examine how teachers are actually implementing the portfolio assessment. Our surveys and content analysis indicate substantial variation in implementation that could affect the validity of comparisons based on the portfolio scores as well as their impact on student learning. Specifically,

- ***There were important differences in the ways teachers used the mathematics portfolios in 1991-92.*** Teachers had different rules regarding authorship, revision, and the selection of tasks. Two-thirds of fourth-grade teachers and 44% of eighth-grade teachers

[3] Changes in teachers' responses to the *Resource Book* illustrate the evolutionary impact of the program on instruction. During the first full year of implementation, many teachers had difficulty finding tasks that were appropriate for portfolios, and the *Resource Book* was seen by many as a valuable source. By the end of the 1992-93 school year, however, teachers' understanding of the types of problems that students should be given had evolved markedly, and many teachers had come to view some of the problems in the *Resource Book* as inadequate. Indeed, at the state scoring session in June, 1993, a number of teachers suggested "recalling" the *Resource Book* because of a perceived inadequacy of the tasks it includes.

set ground rules for the amount or type of assistance students could receive with portfolio pieces; the remainder of the teachers did not. Approximately one-half of the teachers said portfolio pieces were revised one time; about one-quarter of the teachers reported no revisions, and about one-quarter reported two or more revisions.

A review of the contents of portfolios from three fourth-grade classes and three eighth-grade classes found wide variation in the tasks included in portfolios within a given class and even wider variation in portfolio tasks between classes. There were 70 different tasks included in the portfolios of the three fourth-grade classrooms we examined, and there were 90 different tasks in the portfolios of the three eighth-grade classrooms.[4] Together, there were 154 different tasks listed on the tables of contents of the six classrooms (100 portfolios) we examined (see Tables 1 and 2).

Within a classroom, many tasks appeared in only one student's portfolio, while only a few appeared in the majority of the portfolios. The average percentage of portfolios containing each task ranged from 13% to 50%, with most classes falling below 30%. This indicates a moderate amount of variation in the selection of tasks within classrooms. Another measure of within-class variation is the percentage of tasks that appear in one-half of the portfolios. Very few tasks met this benchmark in any of the classrooms, and in no classroom did we find a task that appeared in all of the portfolios (see Tables 1 and 2).

There was even greater variation in task selection between classrooms. Only two tasks were used by more than one classroom in each grade. Perhaps most striking is the fact that more tasks were shared between grade levels (seven) than among classes within grade levels (four; see Table 3).

Despite the great variation in the assignment of portfolio tasks and in rules governing their use, there were also important similarities in the ways teachers used the mathematics portfolios. For example, almost all teachers emphasized student interest and mathematical correctness as the key criteria for selecting

---

[4] Some tasks appear at both grade levels; therefore the overall count of tasks differs from the sum of the grade level counts. Similarly, grade level counts differ from the sum of the classroom counts because some tasks appear in more than one classroom.

Table 1

Vermont Math Portfolio Task Variation in Three Fourth-Grade Classrooms

Grade 4: 43 portfolios, 70 different tasks		
Classroom 4-1		
Number of portfolios		11
Number of tasks per portfolio		6-7
Number of different tasks in classroom		28
Average number of times each task appeared		2.6
Average percent of portfolios containing each task	23%	
Number of tasks appearing in more than 50% of the portfolios		3
Classroom 4-2		
Number of portfolios		14
Number of tasks per portfolio		5-7
Number of different tasks in classroom		26
Average number of times each task appeared		3.0
Average percent of portfolios containing each task	21%	
Number of tasks appearing in more than 50% of the portfolios		0
Classroom 4-3		
Number of portfolios		18
Number of tasks per portfolio		7
Number of different tasks in classroom		18
Average number of times each task appeared		7.0
Average percent of portfolios containing each task	39%	
Number of tasks appearing in more than 50% of the portfolios		4

pieces to go into the portfolios. Similarly, in both fourth and eighth grades, all students in participating classes compiled portfolios.[5]

These differences in mathematics portfolio practices may affect the interpretation of portfolio-based scores and therefore the validity of comparisons based on them. Some have suggested that school- or district-level aggregate portfolio scores be used for accountability purposes, and Vermont has discussed

[5] In addition, 80% of the eighth-grade teachers used portfolios in only one or two classes during the 1991-92 school year.

Table 2

Vermont Math Portfolio Task Variation in Three Eighth-Grade Classrooms

Grade 8: 57 portfolios, 90 different tasks		
Classroom 8-1		
Number of portfolios		**12**
Number of tasks per portfolio		**5-7**
Number of different tasks in classroom		**34**
Average number of times each task appeared		**2.4**
Average percent of portfolios containing each task	**20%**	
Number of tasks appearing in more than 50% of the portfolios		**2**
Classroom 8-2		
Number of portfolios		**33**
Number of tasks per portfolio		**5-7**
Number of different tasks in classroom		**47**
Average number of times each task appeared		**4.2**
Average percent of portfolios containing each task	**13%**	
Number of tasks appearing in more than 50% of the portfolios		**2**
Classroom 8-3		
Number of portfolios		**12**
Number of tasks per portfolio		**5-7**
Number of different tasks in classroom		**12**
Average number of times each task appeared		**6.0**
Average percent of portfolios containing each task	**50%**	
Number of tasks appearing in more than 50% of the portfolios		**6**

using portfolio scores to compare schools or larger units.[6] However, comparisons of outcomes based on dissimilar practices may lead to invalid inferences about differences in student performance. For example, it is reasonable to think that the number of times a portfolio piece is revised will affect its quality, which will, in turn, affect the scores it receives on the state's seven criteria. These differences could cause students with similar levels of proficiency to receive different scores and could obscure real differences in student proficiency.

[6] Vermont has many schools with very small enrollments, and some districts contain only a single school at any given grade level. These small enrollments would undermine comparisons on any measure of student characteristics. Accordingly, a commonly discussed alternative would be to use portfolio scores to compare "Supervisory Unions," administrative units that typically include several schools and local districts.

Table 3

Vermont Math Portfolio Task Variation Between Classrooms

Sample	Number of tasks	Number (%) of tasks in more than one classroom[a]
Overall	154	8 (5.2)
Grade 4	70	2 (2.9)
Classroom 4/200	28	0 (0.0)
Classroom 4/201	26	2 (7.7)
Classroom 4/212	18	2 (11.1)
Grade 8	90	2 (2.2)
Classroom 8/000	34	2 (5.9)
Classroom 8/201	47	1 (2.1)
Classroom 8/203	12	1 (8.3)

[a] Percentages calculated based on number in each row.

Similarly, the choice of tasks students are asked to perform and choose to include in their portfolios will determine the skills they demonstrate, which will affect the scores their portfolios receive.[7] Such differences in task selection could also make portfolio scores a distorted indicator of students' actual performance. In addition, there is some evidence that other aspects of classroom context may affect students' performance on particular tasks. During the pilot year, some of the very same mathematics tasks were listed as "most successful" and "most unsuccessful" by different teachers.

- ***Variation in mathematics portfolio use was less marked in 1991-92 than in the 1990 pilot year.*** **During the pilot year there was much greater between-teacher variation in portfolio practices than those noted above. Some eighth-grade mathematics teachers selected a single class for participation, others used portfolios in all their classes. While the typical portfolio contained 6 pieces of work, the number of pieces ranged from zero to 20. About one-quarter of the**

[7] This task selection problem has two components. First, some tasks may offer less opportunity than others to demonstrate a given skill. Second, some tasks may make it easier to demonstrate a high level of performance with respect to a given skill. Since the pilot year, some mathematics portfolio scorers have stated that some teachers assign tasks that are too easy for their grade levels, apparently to increase students' ability to show the attributes scored.

classes did not bother to select students' "best pieces" at all. In some classes the selection was left up to the student, in others it was made jointly by student and teachers, and in some classes teachers decided. The criteria used to choose best pieces were equally varied, ranging from pieces that made the students feel good to pieces that were likely to score well.

During the first year of statewide implementation there was much more consistency in all these aspects of mathematics portfolio use. While there was still considerable variation in portfolio practices, the state had taken steps to standardize many aspects of portfolio use that had been troublesome the year before.

Implementation Problems

A reform of this scale and scope is bound to encounter some problems of implementation. In the case of the portfolios, the most serious problems were continuing confusion on the part of some teachers about the purpose of the portfolios and the procedures they were supposed to use, the perceived inadequacy of information from the state about the innovation, and the rapid speed of the reform. The latter problems occurred in both writing and mathematics. Specifically,

- ***Teachers and students remained confused at times about the use of the mathematics portfolios.*** **Three-quarters of the teachers were occasionally or frequently confused about what they were supposed to do with portfolios or how they were supposed to do it. About one-half of the fourth-grade teachers and one-third of the eighth-grade teachers reported that students often were confused by portfolio activities, as well.**
- ***Many teachers had difficulty finding appropriate mathematics tasks.*** **Despite the dissemination of the *Resource Book*, more than one-half of the mathematics teachers reported at least occasional difficulty finding tasks appropriate for inclusion in the portfolios. Moreover, discussions with raters during statewide scoring sessions in the spring of 1993 suggest that the true proportion may have been even higher, because they reported that some teachers**

included inappropriate materials (such as drill sheets) in the 1991-92 portfolios.

- ***Poor communication, insufficient information, and the rapid pace of implementation posed problems for many teachers in both writing and mathematics.*** **Principals had the strongest opinions about these issues. Approximately one-third of the principals felt the state administered the program poorly. They complained about unclear expectations, late or contradictory information, or poor communication from the state. Twenty percent of the principals also mentioned the need for more attention to scoring, and a few (10%) complained that the speed with which the assessment was implemented statewide was a source of stress.**

The responding mathematics teachers' comments supported these assertions. Approximately one-third of the teachers who were interviewed felt that their work could have been made much easier by a longer implementation period and wider availability of tasks.

The Vermont State Department of Education has been responsive to past concerns of this sort. For example, pilot year training was widely perceived to be insufficient. Teachers were dissatisfied with state-sponsored workshops because they failed to provide specific guidelines for implementing portfolios, adequate numbers of examples of appropriate activities, and clarification of the criteria to be used to judge the portfolios. Furthermore, once the state-sponsored regional workshops were completed, teachers received little, if any, additional support from colleagues or from state consultants. Many teachers felt isolated and underprepared.

The Department of Education tried to address these concerns in 1991-92 by broadening training, publishing specific guidebooks for teachers, and creating regional portfolio assessment networks. The *Resource Book* and the *Teacher's Guide* appeared to be quite helpful to teachers. In fact, as noted above, the *Resource Book* was teachers' chief source of portfolio tasks. The Network Consultants were another source of information, and those teachers who had contact with the consultants thought their feedback was helpful. It would appear that these efforts were improvements, but that teachers still had unmet needs in the area of training.

Effects

According to reports by teachers and principals, the portfolio assessment program was the impetus for diverse changes in mathematics curriculum and instruction, and it engendered both positive and negative reactions among teachers and students. Data from the questionnaires and interviews provide information about the following questions:

- **What effects did the mathematics portfolios have on the content of the curriculum and the style of mathematics instruction?**
- **In what ways did the mathematics portfolios affect student performance?**
- **What effects did the portfolios (both mathematics and writing) have on teacher and student attitudes, including teachers' judgment of student abilities?**

Changes in Curriculum Content and Instructional Style

The mathematics portfolio assessment had a substantial impact on curriculum content, and it appeared to have some effects on instructional activities and style. Specifically,

- ***The amount of classroom time devoted to problem solving increased.*** **More than three-quarters of the teachers reported spending more time teaching problem-solving strategies than they had prior to the portfolios, and approximately one-half reported spending more time teaching patterns and relationships. Between one-third and one-half of the teachers surveyed reported spending less time on computation.[8]**

There was little change in the amount of time spent on the topic of measurement/geometry (see Table 4).

Similarly, most aspects of problem solving received more attention than they had prior to the math portfolios (see Table 5). Only single-step word problems did not receive more attention from the vast majority of teachers. Teacher interviews confirmed these findings. Increased attention to problem

---

[8] Although not asked directly about areas of reduced emphasis, a small number (approximately 10%) of the teachers interviewed volunteered that they spent less time on computation than previously.

Table 4

Change in Class Time Devoted to Mathematical Topics (Percentage of Teachers)

Topic	Somewhat or much less	About the same	Somewhat or much more
	Grade 4		
Computation and algorithms	49	53	7
Estimation	8	55	37
Patterns/Relationships	5	46	48
Measurement/Geometry	15	57	27
Problem-solving strategies	2	16	82
	Grade 8		
Computation and algorithms	31	66	3
Estimation	3	63	33
Patterns/Relationships	10	40	50
Measurement/Geometry	3	73	23
Problem-solving strategies	0	23	77

Note. **Percentages may not sum to 100 because of rounding errors.**

Table 5

Change in Time on Problem-Solving Activities (Percentage of Teachers)

Topic	Somewhat or much less	About the same	Somewhat or much more
Grade 4			
Exploring mathematical patterns	1	38	51
Single-step word problems	16	64	20
Multiple-step word problems	8	38	54
Logic or reasoning problems	1	23	76
Applying math knowledge to new situations	1	26	73
Collecting, analyzing, reporting data	1	27	71
Grade 8			
Exploring mathematical patterns	3	41	56
Single-step word problems	13	69	19
Multiple-step word problems	0	41	59
Logic or reasoning problems	0	25	75
Applying math knowledge to new situations	0	19	81
Collecting, analyzing, reporting data	0	41	59

Note. **Percentages may not sum to 100 because of rounding errors.**

solving and the use of new kinds of problems were the most frequently mentioned changes in teaching.

- ***Teachers' emphasis on mathematical communication increased.*** **Approximately two-thirds of the teachers reported that they placed greater emphasis on oral discussions of mathematics, making charts and graphs, and writing reports than they had prior to the portfolios (see Table 6). Some of the teachers who were interviewed specifically mentioned spending more time on writing in the area of mathematics.**
- ***The amount of time students worked in small groups and in pairs increased.*** **Approximately one-half of the teachers reported that more mathematics work was done in small groups and in pairs than in prior years. More group work was also mentioned during the teacher interviews. In comparison, most teachers reported no change in the amount of time students spent working individually or in whole-class activities.**

Table 6

Change in Emphasis on Aspects of Mathematical Communication (Percentage of Teachers)

Topic	Somewhat or much less	About the same	Somewhat or much more
Grade 4			
Describing personal experiences	2	47	51
Oral discussions of mathematics	0	31	69
Making charts, graphs, diagrams, etc.	3	28	68
Written reports about mathematics	1	33	66
Grade 8			
Describing personal experiences	3	69	28
Oral discussions of mathematics	6	44	50
Making charts, graphs, diagrams, etc.	0	44	66
Written reports about mathematics	3	25	75

Note. **Percentages may not sum to 100 because of rounding errors.**

- ***Schools voluntarily expanded the use of portfolios to other grade levels.*** **Nearly one-half of the principals indicated that the portfolio program had already been expanded beyond the fourth and eighth grades, and others said they intended to expand it in the future. Many teachers also recommended expanding the program; they commented that for the program to be a success, it must be implemented in all elementary and middle school grades. This occurred despite the considerable burdens that accompanied the use of the portfolios (see below).**

- ***Other mathematics instructional practices appear to have changed to some degree.*** **The majority of all principals interviewed stated that the program has had beneficial effects on mathematics instruction, including in their responses a variety of changes in curriculum content as well as instructional methods or styles. The effects they mentioned were diverse, including increased emphasis on "flexible thinking," lessened reliance on textbooks, less emphasis on drill and practice, increased reliance on hands-on learning, increased use of interdisciplinary projects, and an increased emphasis on communication of mathematics.**

In response to another question, almost one-half of the principals interviewed mentioned the value of the portfolios as an educational intervention. They also reported positive changes in curriculum, better communication and collaboration among teachers, higher levels of thinking and work, a broadening of individuals' views of mathematics and of mathematics activities, a movement away from traditional mathematics (by, among others, teachers who otherwise would not have made those changes), and a lessening of "math phobia."

Finally, one-quarter of the principals stated that it was too early to accurately assess the impact of the portfolio program on mathematics instruction.

Changes in Student Mathematics Performance

We did not attempt to assess changes in student performance directly, but teachers were asked to comment on changes they observed in the work students produced, particularly differences between portfolio tasks and traditional mathematics assignments. Teachers reported considerable change in student

performance, which affected their own opinions about student ability. Specifically,

- ***Regardless of ability level, most students performed differently on portfolio tasks than on regular mathematics assignments.*** **Only about one-third of teachers said students' performance was the same on both types of tasks. About one-quarter of the teachers said performance varied so greatly across tasks or across students that they could not make overall comparisons. Of the remaining teachers, most reported that students did more poorly on portfolio tasks than traditional assignments.**

Similarly mixed results were reported when teachers were asked to focus on low-ability students. Approximately one-quarter of the teachers reported that low-ability students were frequently more successful as a result of the portfolios, and they were "occasionally" more successful in another 40% of the classrooms (see Table 7). On the other hand, virtually all fourth-grade teachers and 80% of the eighth-grade teachers indicated that low-ability students had difficulty with portfolio tasks at least occasionally (see Table 8).

- ***Teachers' judgments of students' mathematics abilities changed as a result of student portfolio work.*** **More than 80% of teachers said they had changed their opinion of students' mathematical ability on the basis of students' portfolio work. Although the amount of change reported by most teachers was small, the pervasiveness of change was striking. Moreover, one-third of teachers said they changed their opinion of students' abilities a "moderate amount," and nearly 10% changed their opinions "a great deal." Pilot year teachers said the portfolios permitted students to demonstrate greater creativity and differentiate themselves one from another more than traditional assignments.**

Changes in Teacher and Student Attitudes

The introduction of the portfolios caused changes in teacher and student attitudes towards curriculum, instruction, and learning. On the positive side, there was increased enthusiasm among mathematics teachers for their subject; on the negative side, both mathematics and writing teachers were frustrated

Table 7

Frequency of Positive Effects (Percentage of Teachers Reporting)

Issue	Rarely or never	Occasion-ally	Often or always
Grade 4			
I am more enthusiastic about teaching math	15	29	56
Goals of math instruction are improved	10	33	57
Math is more closely linked to other subjects	14	41	45
Students' attitudes toward math improve	19	38	43
Students are learning more mathematics	14	35	51
Low ability students are more successful	28	40	33
Grade 8			
I am more enthusiastic about teaching math	11	38	51
Goals of math instruction are improved	14	57	30
Math is more closely linked to other subjects	11	46	43
Students' attitudes toward math improve	32	41	27
Students are learning more mathematics	14	51	35
Low ability students are more successful	38	46	16

Note. **Totals may not sum to 100% due to rounding errors.**

about changes that were often perceived to be demanding and sometimes perceived as threatening. Specifically,

- ***Portfolios increased mathematics teachers' enthusiasm for their subject and had other positive effects on teachers' attitudes.*** **Surveyed mathematics teachers reported that the portfolio program had a number of positive effects on their own feelings about teaching mathematics. For example, more than one-half of the teachers said they were frequently more enthusiastic about teaching math, and over 90% were more enthusiastic at least occasionally. Similarly, over 40% said the goals of mathematics instruction were improved and math was more closely linked to other subjects (see Table 7).**

Principals spoke more generally about the effects of the portfolios (both writing and mathematics) on teachers' attitudes. Although the great majority of principals characterized the attitudes of their teachers to the portfolio program as mixed, 23% mentioned specific positive feelings on the part of teachers. Only

Table 8

Frequency of Portfolio-Related Problems (Percentage of Teachers Reporting)

Issue	Rarely or never	Occasion-ally	Often or always
Grade 4			
I don't understand what I'm expected to do	24	50	26
I don't have enough training in how to do it	25	39	36
I have difficulty finding appropriate tasks	25	42	33
I lack time to prepare portfolio lessons	15	25	59
Not enough time to cover the full math curriculum	4	15	81
Low ability students have difficulty with tasks	3	37	60
Students don't understand what to do with tasks	9	45	45
Students don't know how to solve problems	7	43	50
Students not interested in portfolio tasks	29	50	21
Grade 8			
I don't understand what I'm expected to do	27	57	16
I don't have enough training in how to do it	38	43	19
I have difficulty finding appropriate tasks	41	41	19
I lack time to prepare portfolio lessons	14	22	65
Not enough time to cover the full math curriculum	11	24	65
Low ability students have difficulty with tasks	19	31	50
Students don't understand what to do with tasks	11	50	39
Students don't know how to solve problems	8	57	35
Students not interested in portfolio tasks	16	51	32

Note. **Totals may not sum to 100% due to rounding errors.**

4% of the principals characterized their teachers' attitudes about the program as negative. About one-fifth of the principals (18%) made generic comments about teachers' positive responses to the concept of the portfolio program.

- ***Teachers and principals perceived the portfolio program to be a substantial burden.*** **Eighty-six percent of principals interviewed labeled the portfolio program as "burdensome" (without regard to subject matter), although they appeared to view the burdens as primarily resting on teachers rather than on themselves. (Burdens will be discussed in greater detail in the next section.) For example, 26% of the principals noted that teachers found the time they had to spend on the portfolios to be burdensome. Ten percent mentioned**

that teachers felt they had to be out of the class too often in order to attend the necessary training sessions. Ten percent also stated that teachers felt it was inappropriate to have the fourth-grade teachers burdened by both mathematics and writing portfolio programs. Some of these time pressures probably will decrease as teachers become more familiar with the innovation; others are likely to persist.

- ***Scoring was a particular source of anxiety among teachers.*** **During the pilot year, some teachers expressed concerns about the use of portfolios to judge students and school performance. These concerns continued in 1991-92. Roughly one-third of the fourth-grade mathematics teachers and one-half of the eighth-grade mathematics teachers voiced concerns about the validity and fairness of the scoring system and about the public reporting of scores. This is a significant response given that these were unprompted comments. Principals echoed these concerns more generally. About one-fifth of the principals mentioned problems their mathematics and writing teachers were having with scoring, including concerns about subjectivity, inconsistency, and training.**

- ***Students had mixed reactions to the mathematics portfolios.*** **We obtained no direct measures of students' reactions to the portfolio program, but we did question teachers and principals about them. Over 40% of the fourth-grade mathematics teachers and about 30% of the eighth-grade mathematics teachers said that students' attitudes towards math improved as a result of the portfolios and that students were learning more mathematics (see Table 7). On the other hand, one-half of the teachers at each grade level reported that students strongly disliked writing about mathematics and explaining the thinking that went into their work, both of which were common elements of many portfolio tasks.**

Students' positive reactions to the mathematics portfolios included pride in their work, excitement over new challenges, interest in problems and problem solving, enjoyment of cooperative group work, fun with new types of activities, and enthusiasm for expressing themselves in writing. Students' negative reactions to the mathematics portfolios primarily centered on their dislike of

writing, but also included frustration with new and unfamiliar activities, and, among eighth-grade students, lack of understanding of purpose, anger over the amount of time involved, and occasional concern about the effect on their grades.

Seventy percent of the principals described students' responses to the mathematics portfolios, and the positive responses far outweighed the negative ones. About one-fifth of the principals noted specific aspects of the mathematics portfolios about which students felt positive, including the types of problems they were working on, finding creative solutions, doing hands-on projects, using manipulative aids, using language arts in mathematics activities, and doing interdisciplinary work. Some principals commented that portfolio work engendered feelings of pride or ownership. Principals' comments about negative student reactions to the mathematics portfolios were diverse and difficult to summarize. Only one point was made by more than a single principal: Eight percent of the principals mentioned that students do not like the portfolios because of the writing involved; in particular, they reported, students found it difficult to write down their thought processes and disliked doing so. About 10% of the principals observed that student attitudes toward the portfolios were related to teacher attitudes.

Burdens

Change has come at a price. The portfolio assessment placed substantial new demands on students, teachers and principals. The surveys and interviews provide partial answers to two general questions:

- **What demands did the mathematics portfolios make on teachers' and students' time?**
- **What demands did the mathematics and writing portfolios make on other school resources?**

Time Demands

The mathematics portfolios placed major time demands on teachers.

- ***The mathematics portfolios required a significant amount of class time, which had to be taken from other activities.*** **Eighty percent of the fourth-grade teachers and 65% of the eighth-grade teachers had difficulty finding time to cover the regular mathematics curriculum. On average, fourth-grade classes spent 15 hours each month on the**

math portfolios, and eighth-grade classes spent 10 hours each month (see Table 9). Principals were aware of these efforts, and a large number of principals complained about the burdens of record keeping, logistics, and the demands on class time.

One consequence of the increased time devoted to mathematics portfolios was a reduction in attention to other parts of the curriculum. Over 80% of fourth-grade teachers and over 60% of eighth-grade teachers reported that they often had difficulty covering the required curriculum (see Table 8). In interviews, several fourth-grade teachers, who usually teach in self-contained classrooms, reported that the portfolios had forced them to spend more time on math and less time on other subjects. Several eighth-grade teachers, who usually teach in departmentalized settings, reported they had to cover fewer math topics.[9]

These demands on classroom time are not likely to diminish significantly in the future; they are an essential part of the portfolios. If the portfolio assessment program is to be successful, teachers will need to find ways to accommodate these demands on class time while not sacrificing essential elements of the curriculum.

- ***Mathematics portfolios required a significant amount of teacher preparation time.*** **The portfolio program also demanded a large amount of time outside of the classroom. On average, teachers spent 17 hours each month finding portfolio tasks, preparing**

Table 9

Classroom Time Spent on Portfolio Activities (Hours per Month)

Activity	Grade 4		Grade 8		Overall	
	Mean	(*SD*)	Mean	(*SD*)	Mean	(*SD*)
Doing portfolio tasks for the first time	7.8	(5.8)	5.3	(4.7)	7.1	(5.8)
Revising or rewriting portfolio tasks	4.1	(5.7)	2.2	(3.4)	3.6	(5.2)
Organizing/managing portfolios	3.0	(3.3)	2.3	(2.9)	2.8	(3.2)
Total classroom time	15.0	(13.0)	9.9	(8.9)	13.7	(12.2)

[9] This is consistent with teachers' responses to other types of high-stakes testing programs. (See, for example, Salmon-Cox, 1982, 1984.)

portfolio lessons, and evaluating the contents of portfolios (see Table 10). Sixty percent of the teachers at both grade levels said they often lacked time to prepare portfolio lessons (see Table 8).

Principals also noted that teachers had to find time to review and manage the portfolios. Fourth-grade teachers were especially pressed for time because they had to implement both writing and mathematics portfolios in the same year.

In theory, preparation time should diminish over the next few years as teachers build a repertoire of appropriate tasks and activities that can be reused or adapted. There may also be a decrease in the time teachers spend managing the portfolios, as they develop strategies for handling these responsibilities more efficiently. It is unlikely, however, that these demands will ever disappear.

Resource Demands

Both the mathematics and writing portfolios placed demands on school resources. Specifically,

- *Most principals provided release time to teachers for portfolio training and other activities.* Almost all principals (92%) provided special support to their mathematics and writing teachers participating in the portfolio project, most often release time. About three-quarters of principals provided release time for teachers to attend state-sponsored training workshops, and more than one-fourth of principals provided release time for teachers to work on

Table 10

Teacher Time Spent on Portfolio Activities (Hours per Month)

Activity	Grade 4		Grade 8		Overall	
	Mean	(*SD*)	Mean	(*SD*)	Mean	(*SD*)
Finding appropriate tasks and/or materials	6.0	(6.7)	5.9	(5.2)	5.9	(6.3)
Preparing portfolio lessons	6.3	(5.1)	6.9	(6.8)	6.5	(5.6)
Conducting portfolio lessons	8.8	(5.5)	10.3	(14.5)	9.2	(8.9)
Helping students organize/ manage their portfolios	3.7	(4.5)	2.7	(2.8)	3.4	(4.1)
Scoring/evaluating the contents of portfolios	4.6	(7.1)	6.7	(7.6)	5.2	(7.3)
Total teacher time	28.9	(22.0)	33.2	(27.2)	30.1	(23.6)

preparing lessons, selecting best pieces, organizing final portfolios and other portfolio activities outside of class during the school day. A few principals commented that the time required for training was excessive, and some teachers felt they were being asked to spend too much time away from their classes to attend portfolio training sessions.

The provision of release time was a substantial financial burden for participating schools, because it was common for schools to bear the cost of substitute teachers when release time was granted. Nonetheless, very few principals (13%) reported that they were unable to provide the support teachers requested.

The need for extensive training should decline as teachers become familiar with the basic elements of the portfolio assessment program. Additional state-sponsored training continued during the 1992-93 school year. Beyond that, there will always be a need for some additional training to prepare new teachers, to supplemental teacher expertise in new ways as the assessment reform matures, and to maintain standards in the implementation of the program and scoring of the mathematics and writing portfolios. However, it is likely that training demands will decrease somewhat over the coming years.

Conclusions

During 1990-91 and 1991-92, Vermont made important strides toward realizing the goals of the portfolio assessment program for the reform of assessment and instruction. The innovation was widely adopted, and mathematics teachers reported paying substantially more attention to problem solving and mathematical communication. There was also some evidence of changes in instructional practices, and the state appeared to be promoting greater acceptance and more effective use through improved teacher training. There was widespread support for the reform at the school level; nearly one-half of the schools were voluntarily expanding the use of portfolios to other grade levels.

However, substantial problems remain. From the practitioner's perspective, the mathematics portfolio assessment has created new burdens for principals, teachers and students. These burdens include demands on teachers' time and

school resources for training, preparation and use of portfolios. While some of the demands created by portfolios are likely to decline with experience, others represent continuing burdens.

The variability we found in the implementation of the mathematics portfolio program, while consistent with the notion of a decentralized, bottom-up reform, was so substantial that it threatens both the impact of the program on instruction and the validity of the resulting data. In particular, we are concerned about the apparent lack of a common understanding of the fundamental constructs of problem solving and mathematical communication. If the assessment program is to achieve its goals, the state will need to find ways to instill a shared understanding of these core constructs. In addition, Vermont needs to be concerned about teachers' idiosyncratic portfolio practices, which will affect its ability to interpret portfolio scores.

These observations pertain only to the initial stages of this complex intervention; it is too soon to tell what the ultimate impact of the portfolios will be on curriculum and instruction or how useful they will be for classroom assessment. The answer to these questions will require more time for the program to mature and further study. Similarly, questions about teachers' understanding of problem solving and mathematical communication, the impact of these reforms on students from different backgrounds and students of different abilities, and the long-term acceptance and transformation of the vision of portfolios guiding Vermont's efforts will be answered only in the future.

CHAPTER 3. RELIABILITY OF WRITING PORTFOLIO SCORES

This chapter examines the reliability of scores assigned to writing portfolios. We found that writing scores were unreliable because raters were inconsistent in their evaluations of the quality of students' work. (Chapter 4 examines the reliability of scores assigned to mathematics portfolios, and Chapter 5 reports on the reliability of school-level portfolio scores.)

"Reliability" has many meanings, however, so it is important to clarify which aspects of reliability we assessed. For the most part, we examined "rater reliability," which is the degree to which different raters agreed on the scores that should be assigned to portfolios or to individual pieces within them. Rater reliability is critical, because it limits the amount of confidence that can be placed in the scores. However, it is only one aspect of the larger question of the reliability (or consistency) of scores, which is often called "score reliability." For scores to be reliable in the broader sense, they must be consistent, not only across raters, but also across different instances of measurement. For example, a score is not a reliable measure of a student's writing if it is specific to the particular essays the student is asked to write for the test. Comparison to a multiple-choice test clarifies this distinction. The "rater reliability" of a multiple-choice test is perfect; the scanning machine and software will generate exactly the same score from a given answer sheet time after time. When publishers of multiple-choice tests report less than perfect reliability, they are referring to differences in students' scores on different forms of the test or on halves of the same test.

High rater reliability is also necessary but not sufficient to indicate that scores are an accurate reflection of the quality of work on the specific tasks students have included in their portfolios. If two raters do agree on the scores that should be assigned to a portfolio, it could be because they both made accurate assessments of that portfolio's quality. However, a high degree of agreement between raters also could occur for reasons that have little or nothing to do with portfolio quality, for example, if they gave high marks to portfolios that were written neatly and low marks to messy ones. Low agreement rates

necessarily indicate that at least one of the raters did not make an accurate assessment of portfolio quality.

This chapter examines the following specific questions about the reliability of ratings of the 1992 Vermont writing portfolios:

- **Did raters agree with each other regarding the relative quality of the portfolios?**

- **Was there more agreement among raters on some grading dimensions (often called "scoring criteria" in Vermont) than on other dimensions? Is there more agreement between raters regarding what scores should assigned to the best piece in a portfolio than there is on what scores should be given to the combination of other pieces in that portfolio?**

- **Were some raters generally more lenient than others, and does a student's own teacher generally assign a higher or lower grade to a portfolio than an independent rater?**

- **Did students who received relatively high scores on one dimension also tend to receive high scores on other dimensions?**

- **Did students who received a high score on their "best" piece also tend to receive a high score on the remainder of their portfolio; that is, is a student's performance consistent across different parts of the portfolio?**

- **To what degree would the reliability of portfolio scores be improved by increasing the number of independent raters who evaluate a portfolio or increasing the number of separately graded parts?**

- **Are the answers to the questions above the same for Grade 4 as they are for Grade 8? For instance, were Grade 4 portfolios scored any more or less consistently than Grade 8 portfolios?**

- **How did the degree of agreement between raters in grading portfolios compare to the agreement rate in grading essays when all students wrote responses to the same prompt under standardized conditions?**

Chapter Overview

The next portion of this chapter describes the procedures that were used to gather information about the degree of agreement between raters. We then show the typical agreement rate on a dimension. This is followed by a summary of agreement rates on all dimensions at both grade levels, an analysis of some of

the factors that may have contributed to these rates, a discussion of the effects of different strategies for improving the reliability of portfolio scores, and a comparison of rater agreement rates on portfolios with those on a uniform writing test. Our conclusions about rater agreement appear at the end of the chapter.

Procedures

As noted previously, each writing portfolio had two parts. One part was the piece the student and/or teacher selected as the "best" one. The other part was the remaining 5 to 7 pieces taken together as a set. These two parts are referred to as the "best" and "rest," respectively.

Both parts were graded on five dimensions by the classroom teacher: purpose, organization, details, voice and tone, and the combination of usage, grammar, and mechanics (see Appendix D for a description of each dimension). Thus, all told, each rater assigned 10 scores to a portfolio (5 to the best piece and 5 to the rest). Each of the 10 scores was assigned on a 4-point scale. A rater assigned all 10 scores before going on to the next portfolio.

A sample of 1,903 Grade 4 portfolios and 750 Grade 8 portfolios were selected randomly from among those that were already scored by the student's own classroom teacher. Each selected portfolio was graded again by one of the teachers who participated at a centralized portfolio grading workshop. This second rater did not see the scores assigned by the first rater, that is, the student's own teacher. All of the second raters were drawn from the pool of Vermont classroom teachers, but none of them regraded portfolios from their own classroom.

Agreement Between Raters

We addressed two primary questions about the degree of agreement between the scores assigned to portfolios by the student's own classroom teacher and those assigned by an independent rater. First, were the scores assigned by classroom teachers biased, that is, systematically higher or lower? Second, to what degree were raters consistent with each other in deciding which portfolios warranted high and low scores?

Bias

Some observers expressed concern that students' own teachers might be more lenient in assigning scores. That did not happen. On average, a student's teacher was only trivially more lenient than was the second rater. Averaged across dimensions, the scores provided by the students' own teachers were never more than 0.04 points (on the 4-point scale) higher than those assigned by the second raters (Table 11). This striking similarity of average ratings was also quite consistent across specific scoring dimensions. In only 4 of the 20 comparisons (2 grades x 2 parts x 5 dimensions) was the mean difference between raters greater than 0.05, and in no case did it exceed 0.10 (see Appendix A).

Consistency of Scores

Even though the two raters assigned very similar *average* scores, the reliability of ratings was low: Raters were not consistent in assigning scores to individual portfolios.[1] That is, although they agreed on the average score across all portfolios, raters often disagreed about which portfolios warranted high or low scores. The small differences in mean scores noted above contributed virtually nothing to this inconsistency in ratings.

Table 11

Mean Scores Across Dimensions by Rater Type, Grade, and Part

	Grade 4		Grade 8	
Rater type	Best	Rest	Best	Rest
Classroom Teacher	2.86	2.70	2.99	2.77
Independent Rater	2.82	2.66	2.96	2.77
Difference	0.04	0.04	0.03	0.00

Note. **Despite the large number of cases, these differences are not statistically significant. The t-values for best and rest at Grade 4 are 1.47 and 1.30, respectively. The corresponding values at Grade 8 are 0.90 and 0.02.**

[1] **More precisely, they were inconsistent in assigning scores to the two parts of the portfolio, the best piece and the rest. Although we combined scores across the two parts and discuss the reliability of the combined scores later in this chapter, the best-piece and rest scores were not combined into an overall portfolio score by the assessment program.**

This section discusses the evidence for this conclusion. We begin by illustrating the seriousness of the reliability problem by showing the degree of agreement on one scoring dimension for one part of the portfolio. We then present summary data to show that the very low level of agreement between raters in this slice of the data is typical of that found on both parts of a portfolio, on all five dimensions, and at both grade levels.

We use two different approaches to summarize the reliability of ratings. One approach examines the proportion of cases in which both raters agree on a score. This percentage agreement must be interpreted cautiously, however, because some amount of agreement would be expected by chance alone. (The agreement expected by chance is what one would obtain if portfolios were scrambled randomly, so that one would be comparing the first rater's score for one student with the second rater's score for a randomly chosen student, who would in almost all cases be a different student.)[2] The rate of agreement expected by chance varies depending on the number of score categories and the degree to which scores are concentrated in one or a few categories. When scores are highly concentrated at a single value, the rate of agreement expected by chance can approach 100%. (This happened in mathematics but not in writing; see Chapter 4.) A second conventional measure of rater reliability is the correlation coefficient between raters. The correlation coefficient ranges from a value of 0.00 (if there is no relationship between the scores provided by two raters) to 1.00 (if there is a perfect relationship).[3]

---

[2] The chance rate is a function of each reader's distribution of grades across the four score levels. If both raters assigned 25% of the portfolios to each of the four possible scores—a highly unlikely case—then one would obtain agreement in 25% of the cases by chance alone. More generally, however, the chance rate of agreement depends on the "marginal percentages"—i.e., the percentages of portfolios assigned to each score. The chance rate is the sum of the products of the marginal percentages for each rater. For example, suppose that 33% of first raters and 32% of second raters assigned a value of 1. The product of those percentages is .33 x .32 = .11. Products are similarly calculated for scores of 2, 3, and 4, and the four products are summed, and that sum is the percentage agreement expected by chance.

[3] We used the Spearman rank order correlation (rho) rather than the more common Pearson correlation because the difference between a 1 or 2 on the 4-point grading scale was not always viewed the same as the difference between a 2 and 3 or between a 3 and 4. In practice, however, the choice between the Spearman and Pearson coefficients made little difference. We did not use the Kappa statistic to quantify agreement rates because it is not sensitive to distances off the diagonal. Kappa treats a 1-point difference between readers the same as a 3-point difference between them.

The Typical Pattern

Regardless of which index was used (percentage agreement or correlation coefficient), the degree of agreement between raters was low. It was generally similar on all five dimensions on both parts and was only slightly higher at Grade 8 than at Grade 4.

Table 12 shows the extent to which different raters assigned the same or different scores to the best piece on the Details dimension at Grade 4. The rows correspond to the possible scores assigned by the first rater (i.e., the student's own classroom teacher). The columns refer to the possible scores assigned by the independent rater at the special grading session. The entries in the table are the percentages of twice-scored portfolios receiving each of the 16 possible combinations of scores. For example, 1% of the portfolios received a score of 1 from both raters. However, in 2% of the portfolios, the first rater assigned a score of 1 and the second rater assigned a 2. On another 1% of the portfolios, the first rater assigned a score of 1 and the second rater assigned a score of 3.

If the first and second rater in Table 11 agreed perfectly with each other on how each portfolio should be scored, then the values along the diagonal (from upper left to lower right) would sum to 100%. Instead, raters agreed with each other on only 45% of the portfolios. (The sum of the diagonal entries is 1 + 16 + 21 + 7 = 45.) By chance alone, two raters would agree on the score assigned to

Table 12

Rater Agreement—Typical Pattern (Best Piece-Details-Grade 4 Writing)

	Second rater				
First rater	1	2	3	4	Total
1	1	2	1	0	4
2	2	16	12	3	33
3	2	12	21	9	44
4	0	3	10	7	20
Total	5	33	44	19	100%

Note. **Percentage of pieces on which the two raters agreed on what score to assign = 1 + 16 + 21 + 7 = 45%. Percent by chance alone = 36%. Kappa = 0.17. Spearman's rho = 0.35.**

roughly 35% of the portfolios on this dimension. Thus, the observed degree of agreement, 45%, is only about 10 percentage points better than chance.[4]

The correlation coefficients between the grades assigned by the classroom teacher and the independent rater also showed that rater reliability was generally low (Table 13). The correlations varied from a low of .28 (Voice/Tone for the "rest" in fourth grade) to .57 (Usage, Grammar, and Mechanics for the rest in eighth grade). Only one of the 20 correlations exceeded .50, however, and the mean correlations (averaged across dimensions) were only .35 in fourth grade and .43 in eighth grade. These are low by any standard. A correlation of .35 means that one can predict about only 12% in the variance of second raters' scores by knowing the first raters' scores; a correlation of .43 means that one can predict about 18%. The remainder of the variance is error.

Relationships Between Scores on the Two Parts

The Vermont assessment program was designed to provide two separate sets of scores for each portfolio, one for the best piece and another for the rest. By design, no composite score across the two parts was created. Our understanding is that this design had a number of rationales. Keeping the two

Table 13

Interrater Correlations, Writing

	Best		Rest	
Dimension	Grade 4	Grade 8	Grade 4	Grade 8
Purpose	.33	.34	.33	.39
Organization	.36	.41	.31	.43
Details	.35	.44	.33	.41
Voice/Tone	.33	.45	.28	.37
U/G/M	.40	.48	.43	.57
Mean	.35	.42	.34	.43

[4] Some participants in the program maintained that adjacent scores (e.g., a 2 and a 3) should count as agreement. We rejected this argument because the program uses only a 4-point scale. With only four points on the scale, counting adjacent scores as agreement would mean that most scores count as agreement. For example, in the case of Table 12, if the first rater gave a piece a score of 3 (the modal score), then any score assigned by the second reader other than the very rare value of 1 would count as agreement.

parts separate might be desirable in terms of incentives to students and teachers; that is, they might do different types of work in preparing the two parts. In addition, the two parts were supposed to provide different views of a student's performance.

To what degree were the parts of the portfolio actually independent of each other? We cannot ascertain whether the two parts provided different incentives in the classroom. However, we can assess the extent to which the two parts functioned independently as measurement tools. In this section, we compare mean scores on the two parts and examine the correlations between them. In a later section, we use a technique called generalizability analysis to revisit this question. If the mean scores on the two parts are very similar and if the correlations between scores on the two parts are high, then having both parts contributes very little additional, independent information. In other words, in terms of measurement, including the second part does little to improve the information yielded by the assessment.

Overall, our analyses showed that scores on the two parts were strongly related and that the inclusion of a second part added little additional, independent information. This indicates that as of 1991-92, the program was largely unsuccessful in meeting its goal of using the two parts to garner different information about student performance.

Mean Scores

The mean score on the best piece was usually only slightly higher than the mean on the rest (Table 11). This pattern held at both grade levels and for all of the scoring dimensions. Because the best piece is by definition supposed to be better than others, these small differences between parts may indicate that the system is not functioning properly. One possibility is that the students are not successfully identifying a best piece; that is, their best pieces are not much better than any of their other pieces. A second possibility is that the rating system is not sensitive enough to discern differences in quality between best pieces and others. Yet another possibility is that raters have relative standards—that is, they implicitly grade best pieces harder than others. This would undermine the state's effort to interpret scores in terms of the stated definitions of each scale point, which are the same for both parts.

Correlations Between Parts

Some of the simple correlations between scores on the two parts are low. However, this appears to be the result of rater unreliability. When the correlations between parts are adjusted for rater reliability, they are quite high, suggesting that there is relatively little independent information added by including both parts.

Averaged across dimensions, the simple correlations between raters across parts range from .25 to .37 (the "raw" correlations in Table 14). The correlations across parts are somewhat higher in the eighth grade than in the fourth grade. Moreover, it makes little difference whether scores are restricted to a single dimension. In fourth grade, for example, if one compares the score on one part to the score on the other, both on the same dimension, the correlation between the scores is .28. If one compares the score on one part and one dimension to the score on the other part *on another dimension*, the score is only trivially lower: .25.

When these correlations are "disattenuated," however—that is, adjusted to remove the effect of the low agreement rate among raters—they become quite high. The average disattenuated correlation between parts on a single dimension was .77 in the fourth grade and .89 in the eighth grade.[5] In other

Table 14

Mean Correlations Between Parts When Scores Are Assigned by Different Raters, Vermont Portfolios and ITBS Standardization

	Vermont		ITBS, disattenuated	
	Raw	Disattenuated	Same genre	Different genre
Grade 4				
Same dimension	.28	.77	.44	.38
Different dimensions	.25	.73		
Grade 8				
Same dimension	.37	.89	.52	.28
Different dimensions	.34	.81		

[5] These correlations were disattenuated by dividing the raw correlations by the square root of the product of the relevant correlations within piece and dimension but between raters.

words, once one removes the effect of the generally low reliability, scores on one part are fairly strong predictors of scores on the second part, particularly in the eighth grade.

These disattenuated correlations are surprisingly high in the light of experience with other direct assessments of writing, which typically show that ratings of student essays vary substantially within genres and markedly across genres (e.g., Dunbar, Koretz, & Hoover, 1991). For example, the Iowa Tests of Basic Skills include an optional direct test of writing scored with a "focused holistic" rubric. When pieces were graded by different raters, the correlation of scores across different pieces within one mode of discourse averaged only .44 in fourth grade and .52 in eighth grade (after disattenuating for rater unreliability) in a standardization sample.[6] Across genres, the disattenuated correlations were .38 and .28, respectively (Hieronymus, Hoover, Cantor, & Oberly, 1987). For comparison, these correlations are entered in the "different parts, same dimension" rows of Table 24. The Vermont portfolio program maintains only loose control over the content of portfolios, but the guidelines clearly call for the inclusion of different genres, so the lower of these ITBS correlations (between genres) are probably the more reasonable comparison.[7]

A possible clue to the high disattenuated correlations between parts in Vermont can be found in the correlations between the part scores provided by a single rater. Although the raw correlations between parts scored by different raters were very low (before correcting for reliability), the correlations within a single rater tended to be considerably higher. For example, for a single dimension, correlations between parts scored by different raters were .28 and .37 in the fourth and eighth grades, respectively, but they were about .60 when scores were provided by a single rater. (See Table 15; the between-rater correlations are repeated from Table 14 for comparison.) That is, a given rater's score on one dimension for one part of a given portfolio predicts to a moderate degree the score that rater gave the other part on the same dimension.

---

[6] In the case of the ITBS, disattenuation was modest, because interrater reliabilities were very high, ranging from .88 to .99 depending on the grade level and the mode of discourse required by the prompt.

[7] The ITBS scores reflect a single focused holistic score for each piece, which would tend to make them more reliable than the dimension-specific Vermont scores. This makes the higher disattenuated correlations from the Vermont program even more striking.

Table 15

Mean Correlations Between Parts When Scores Are Assigned by the Same and Different Rater

	Same rater	Different rater
Grade 4		
Same dimension	.59	.28
Different dimensions	.44	.25
Grade 8		
Same dimension	.62	.37
Different dimensions	.49	.34

There are several possible explanations for the relatively high degree of consistency across parts within a single rater compared to the much lower consistency across raters. One possibility is that some raters have reasonably consistent rules for rating the portfolios but that different raters have different rules. For example, one rater might be more influenced by inappropriate use of commas than another.[8] This implies either insufficiently precise rubrics or insufficient training in their use. Another possibility is a halo effect—that is, a consistency of scores across parts that is imposed by the rater and that goes beyond the "true" consistency of performance on the parts of a student's portfolio. That is, one or more pieces in a portfolio may color some raters' evaluations of other pieces in the portfolio, so that scores on the later pieces resemble scores on the earlier pieces more than they should.[9] These possibilities are not mutually exclusive.

---

[8] In June of 1993, we observed many hours of discussions by raters working on "calibration pieces" that were used to increase the similarity of their ratings. There were a number of discussions about scoring discrepancies that fit this hypothesis—for example, raters disagreeing about how heavily to weight repeated "usage, grammar, and mechanics" errors of the same type within a single piece. We have no comparable observations of the 1992 scoring that yielded the data presented here, but it is reasonable to assume that such discrepancies were at least as large at that time.

[9] For example, one recent study of writing portfolios from a single elementary school found that scores assigned by teachers to entire portfolios tended to be overly influenced by the score assigned the highest-scoring piece in each (Herman, Gearhart, & Baker, 1993).

Relationships Between Scores on Different Dimensions

A related issue is the degree to which the different scoring dimensions (five in writing, seven in mathematics) function independently. The use of multiple dimensions, like the inclusion of more than one part, has several rationales. Regardless of the patterns shown by scores, employing multiple dimensions may be an effective way to focus instruction on desired attributes of student work. From the perspective of measurement, however, the value of multiple dimensions depends on the degree to which scores on different dimensions provide additional, independent information about the quality of student work. If dimensions do not provide independent information, little would be lost—in terms of measurement—by employing fewer dimensions or by combining scores across dimensions before reporting them.

In general, the scores assigned by a single rater showed considerable consistency across the scoring dimensions. For example, in fourth grade, raters assigned 58% of best pieces exactly the same scores on the Details and Organization dimensions (Table 16). When scores were assigned by different raters, the rate of agreement was much lower: only 40% (Table 17). When a single rater assigned the scores, only 2% of the students had a 2- or 3-point difference between their scores on Details and Organization; when different raters assigned the scores, 10% of the students' scores differed by that much.

Table 16

Degree of Agreement Between the Classroom Teacher's Scores on the Detail and Organization Dimensions (Tabled values are the percentage of students with each of the possible combinations of scores on Organization and Details)

		Organization			
		1	2	3	4
Details	1	1	2	0	0
	2	2	16	14	1
	3	0	8	27	8
	4	0	1	6	14

Note. **Percent of cases with exact agreement: (1 + 16 + 27 + 14 = 58%).**

Table 17

Degree of Agreement Between the Classroom Teacher's Details Score and the Independent Rater's Organization Score (Tabled values are the percentage of students with each of the possible combinations of scores on Organization and Details)

		Organization (2nd Rater)			
		1	2	3	4
Details (Teacher)	1	1	1	1	0
	2	2	11	15	5
	3	1	11	21	10
	4	0	3	10	7

Note. **Percent of cases with exact agreement: (1 + 11 + 21 + 7 = 40%).**

The degree to which raters assign the same scores across dimensions is even clearer when correlations are used to describe agreement. The correlations of scores across dimensions are very similar to the correlations across parts discussed earlier. For example, in fourth grade, the average correlation between the scores on two dimensions assigned by one rater to one part was .57 (Table 18). By way of comparison, the correlation between the scores assigned to the two parts on one dimension by one rater averaged .59. (The correlations between parts in Table 18 are repeated from Table 15 for comparison.) The corresponding agreement rates when scores were compared *across* two raters were much lower (less than .30 in fourth grade), but again the correlations across dimensions were nearly identical to the correlations across parts. (Eighth

Table 18

Mean Correlations Among Dimensions When Scores Are Assigned by the Same and Different Rater, Grade 4

	Same rater	Different rater
Same part		
Same dimension	(1.00)	.34
Different dimensions	.57	.29
Different parts		
Same dimension	.59	.28
Different dimensions	.44	.25

grade results, omitted from Table 18 for simplicity, were very similar, except that the correlations between different raters were about 0.10 higher than the corresponding correlations in the fourth grade.)

One would expect agreement to be lower when scores are provided by two different raters because of the unreliability of ratings documented above; the key question is *how much* lower. To gauge the relative size of these correlations, one can ask: What set of scores provides the best prediction of another? For example, say that one wanted to predict the scores assigned by one rater to one part on a specific dimension. Ideally, the best predictor would be the scores assigned to the same part on the same dimension by another rater; that would indicate that the quality of performance on that part and dimension, rather than idiosyncrasies of raters, determined scores. That was far from the case in Vermont. In the fourth grade, the correlation of scores for the same part and dimension but across raters was only .34, meaning that on average, scores from one rater predict only about 12% of the variance in scores given to that part on that dimension by the other rater ($.34^2 = .12$). In contrast, the scores assigned to that part by the *same* rater but on a *different* dimension predict about 32% of the variance in scores ($.57^2 = .32$). Indeed, it would even be better to know how the same rater scored *the other part on a different dimension* ($.44^2$ = 19% of the variance) than to know how a different rater scored the same part on the same dimension.

When adjustments are made for the low rate of interrater agreement, the correlations between dimensions become extremely high. The most reliable correlations between dimensions that we can obtain are those based on the sum of the two raters' scores for each twice-scored portfolio. On the average, the correlation between two dimensions was about .69 when the score on a dimension was the sum of the two raters' scores on that dimension across the two parts (Table 19). Taken at face value, these correlations would suggest a moderate degree of independence between dimensions, but much of that apparent independence is simply an artifact of random rating error. Virtually all the correlations soar to the middle or upper .90s (close to perfect) when they are corrected for the less than perfect agreement among raters (Table 20). It is therefore likely that any difference between a student's scores on two dimensions results from random error and is not meaningful or interpretable. These findings suggest that in 1991-92, it would have been nearly as useful to assign

Table 19

Interdimension Correlations: Writing, Unadjusted

	Purpose	Organization	Details	Voice/Tone	U/G/M
Purpose		.84	.78	.74	.60
Organization	.84		.75	.69	.67
Details	.81	.79		.76	.57
Voice/Tone	.74	.73	.78		.53
U/G/M	.73	.77	.70	.64	

Note. **Data for Grades 4 and 8 appear below and above the main diagonal, respectively. U/G/M = Usage/Grammar/Mechanics.**

Table 20

Interdimension Correlations: Writing, Adjusted

	Purpose	Organization	Details	Voice/Tone	U/G/M
Purpose		*	*	*	*
Organization	*		*	*	*
Details	*	*		*	*
Voice/Tone	*	*	*		.94
U/G/M	*	*	*	.92	

Note. **Data for Grades 4 and 8 appear below and above the main diagonal, respectively. U/G/M = Usage/Grammar/Mechanics.**

*** Adjusted estimate of correlation is greater than 1.00**

only a single score to each part of the portfolio because the scoring did not reliably distinguish among the different dimensions.

Sources of Variation in Scores: A Generalizability Analysis

We used a statistical technique called "generalizability" analysis (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to investigate the sources of variation in the scores given to students' portfolios. We estimated the degree to which the variation in scores could be attributed to three factors (students, raters, and part) and interactions among them. (See Appendix B for a brief discussion of generalizability analysis and more detail on the analyses summarized here.)

Consistent with the results reported above, the generalizability analysis found that much of the total variation in scores can be attributed to rater error. More specifically, over one-half of the total variance in scores was attributable to nonsystematic differences between raters. This appears as the combination of Rater x Student, Rater x Part, and residual variance components in Table 21.

Only a small part of the remaining variance in scores (about 13% in Grade 4 and about 8% in Grade 8) can be attributed to the tendency for some raters to be systematically more lenient than others (the Rater effect in Table 21). As noted earlier, these small differences in leniency were largely unrelated to whether the rater was the student's own classroom teacher.

Consistent differences among students or between parts accounted for less than half of the total variance in scores. Only about a third of the total variance in scores (28% in the fourth grade and 36% in the eighth grade) can be attributed to differences among students (Table 21). An additional 7% of the total variance in scores was due to the raters agreeing that the student did better on one part than on the other part (the Student x Part interaction). This is consistent with the finding reported earlier that there was hardly any difference in mean scores between parts (Table 11).

This analysis thus provides another index of the severity of unreliability: The majority of the variance of scores arose because of disagreements among raters in their evaluations of the quality of portfolios. Moreover, the analysis further suggests that as of 1991-92, the two parts of the writing portfolio were

Table 21

Percentage of Variance in Writing Portfolio Scores on a Typical Dimension That Was Attributable to Various Factors

Source	Grade 4	Grade 8
Students	28	36
Students x Parts	7	7
Raters	13	8
Raters x Students	18	18
Raters x Parts	1	1
Residual	33	31
TOTAL	100%	100%

for the most part not functioning independently, and little additional information was gained by including both parts rather than one.

Strategies for Improving Score Reliability

The analyses described above found that the score assigned to a part on a given dimension cannot be trusted to provide an accurate measure of that part's quality on that dimension. This problem stems from the very low agreement rates between raters on what score to assign. This section discusses a number of ways to improve reliability: combining scores across dimensions, parts, and raters and adding additional parts or raters. (There are, of course, other strategies that might be used to increase reliability, such as better training and calibration of raters.)

Effect of Averaging Scores Across Parts and Dimensions

More reliable scores for a student can be obtained by combining scores across parts or dimensions. Both types of combining in theory contribute additional, independent information about performance, which should make scores more reliable.

However, in the 1991-92 writing data, even combining across both raters and dimensions increased reliability only modestly. The resulting scores were still too unreliable to warrant confidence that they provide an accurate index of the quality of a given student's work. The reason that the improvement was modest was the sizable correlations between parts and dimensions (within a rater) discussed earlier. That is, because the parts and dimensions were substantially correlated, they provided only limited independent information, so combining them did not greatly improve reliability. For example, Table 22 shows that the average correlation between two raters on one dimension on one part at Grade 4 was only .34. The correlation between raters increases to .39 on a single dimension if the scores on the two parts are averaged, that is, if a student's score on a dimension is the mean of his or her best and rest scores on this dimension. Combining over all dimensions and parts increases the correlation to .49. The same pattern was obtained at Grade 8, but the correlations were all slightly higher. There is little statistical disadvantage to combining scores across dimensions and parts because, as noted earlier, the separate dimension scores are not really interpretable and only a very small

Table 22

Average Correlation Between Raters With Different Types of Combining, Writing

Type of score	4th	8th
No combining—one dimension on one part	.34	.43
One dimension—mean over both parts	.39	.49
One part—mean over all 5 dimensions	.45	.56
Total—mean over all dimensions and parts	.49	.60

portion of the differences in scores between students stems from some of them doing better on one part than on the other part. However, combining in this fashion would eliminate the as-yet unrealized potential for obtaining information about different aspects of a student's work.

Increasing the Reliability of Total Scores by Adding Parts or Raters

If a total score is to be computed for each portfolio, its reliability will be influenced by the number of parts included in the portfolio: Adding more parts will make the total score more reliable. Similarly, combining scores across additional raters will boost reliability. This section discusses the effects of adding parts and raters.

In this context, reliability is the correlation that would be found between different portfolios produced by the same student. That is, each portfolio is considered to be a limited sample of the student's work, and each portfolio is treated as only one of many portfolios that could have been constructed for that student. We have only a single portfolio for each student, but we estimated the correlation that would have been found among different portfolios from the same student by applying standard statistical methods to the scores assigned by the raters (see Appendix B). The resulting estimate is expressed as a correlation (reliability) coefficient. The higher the coefficient, up to a maximum of 1.00, the stronger the estimated relationship between the scores that would be assigned to different, but representative, samples of the student's work.

Increasing the number of raters who evaluate a portfolio from 1 to 2 has a small but noticeable effect on the reliability of total portfolio scores (as defined above) regardless of the number of parts (pieces in the portfolio) that are

evaluated separately. However, there is not much to be gained by having more than 2 raters. Figure 1 shows the estimated relationships between number of raters per portfolio, number of parts evaluated, and reliability for one of the writing dimensions. (Similar figures for all of the writing dimensions are included in Appendix B.) As one benchmark for interpreting these data, a measure should have a reliability of .90 or higher before scores on it are used to make important decisions about individual students (as distinct from larger units, such as schools). Most standardized achievement tests have reliabilities that satisfy this criterion.

Increasing the number of parts within a portfolio that are evaluated improves reliability, but with rapidly diminishing returns. There is not much to be gained by having each rater evaluate more than 5 parts, largely because of the strong correlations among parts discussed previously. This trend holds for any given number of raters. Put another way, it would be difficult to justify having more than two independent raters per portfolio and having them assign scores to more than five separate pieces in that portfolio. However, even under these conditions, reliability is still quite low (.65 to .70) for the purposes of

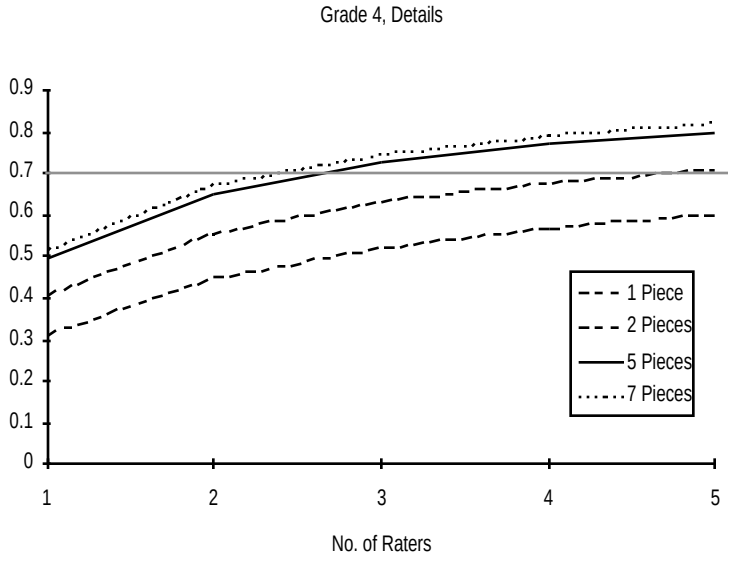


Figure 1. **Effects on reliability of increasing the number of raters or pieces; Grade 4, details.**

making decisions about individual students, particularly given the amount of rater time that would be required to achieve this level (a teacher evaluates only 2 to 3 portfolios per hour).

Comparison to the Uniform Test

In addition to preparing their portfolios, fourth- and eighth-grade students also took an on-demand writing test. The degree of agreement among raters in grading the responses to this test was much higher than it was among those who graded the students' portfolios. This section discusses these differences in reliability and speculates on some of the reasons for them.

The "uniform" exam had a single essay question at each grade level and all the students took this test under the same standardized conditions. Student responses were scored by raters other than Vermont teachers who were hired by a private contractor. These raters used the same five dimensions as were used to grade the portfolios. A sample of 695 fourth graders and 608 eighth graders had their answers read twice. And as with the portfolios, the second rater did not know the score assigned by the first rater.

Interrater agreement levels on the Uniform Test were much higher than they were on the portfolios. The mean correlations at Grades 4 and 8 for the total score across all five dimensions were .87 and .82, respectively. In contrast, the mean correlation between two raters in total portfolio scores (i.e., across all dimensions and parts) was only .49 at Grade 4 and .60 at Grade 8 (see Table 22).

Why were the uniform raters so much more consistent with each other than were the portfolio graders? Each portfolio had 6 to 8 samples of the student's work compared to only a single essay answer on the Uniform Test. Thus, agreement should have been higher on the portfolios because there was more opportunity for chance factors to be balanced out. The students who took both measures came from essentially the same populations, so this was not a source of the disparity either.

A more likely explanation is that high agreement rates can only be obtained when all students respond to the same or similar prompts or when they all produce works that fall within certain well-defined genres, for instance, each portfolio contains one poem, one short story, etc. That did not happen in Vermont in 1992. As a result, portfolio raters were asked to assess whether one

student's response to one task was better than another student's response to a totally different task. This job would challenge even the most conscientious and skilled grader, and given the results discussed above, we now have begun to question whether it can be done with an adequate level of consistency in an operational program.

Conclusions

The findings above led to the following conclusions regarding the reliability of writing portfolio scores in both Grade 4 and Grade 8:

1. **Most raters appear to form a general impression of a portfolio's quality, and this impression seems to affect the scores they assign to all dimensions and parts. However, different raters often develop very different impressions of a portfolio's quality. Consequently, there is a moderate to high degree of consistency across parts and dimensions in the scores a given rater assigns, but a very low degree of agreement between raters in their evaluation of a portfolio's quality.**
2. **The foregoing situation has several ramifications. For instance, any difference in scores among dimensions within a student's portfolio is more likely to be due to chance than to any true difference in student proficiency levels. Thus, any differences in scores between dimensions in a portfolio are not meaningful. The same is true for differences in the scores assigned to the different parts of the portfolios.**
3. **On average, a student's own teacher was only very trivially (and not statistically significantly) more lenient than was the independent rater. Thus, this potential bias was not an important influence on scores.**
4. **On the average, the score assigned to the best piece was only slightly higher than the score assigned to the rest of the portfolio. This pattern suggests that in terms of measurement, there was no functional difference between the best and the rest, and there was nothing gained (in terms of measurement) from identifying one piece as the "best" one.**

5. **Similarly, scores were substantially correlated across dimensions, and therefore including scores on the five dimensions added relatively little information about students' performance.**

6. **Using two independent raters per portfolio and separately grading as many as 5 different pieces in a portfolio will probably improve the reliability of total scores, but increasing the number of raters or pieces yet further would probably produce only minor improvements. Moreover, even basing scores on two ratings of five pieces would not produce a sufficiently reliable score for making decisions about individual students unless rater agreement is increased markedly.**

7. **Scores on the Uniform Test showed much higher rater reliability than did portfolio scores. We suspect that a major reason for this difference is the fact that the Uniform Test required all students to responded to the same question (so that raters could be trained specifically to score responses to that prompt). However, further investigation is needed to separate the effects of standardization of tasks, training differences, and other factors on the reliability of portfolio scores.**

CHAPTER 4. RELIABILITY OF MATHEMATICS PORTFOLIO SCORES

This chapter reports on the reliability of the scores assigned to mathematics portfolios. As in the previous chapter, we focus primarily on the degree of agreement between raters, but we also examine the extent to which the quality of a student's work was consistent across dimensions and across the different pieces in the portfolio.

We found that the raters who graded the mathematics portfolios often disagreed with each other about which score should be assigned. In addition, there was usually less consistency among the pieces in a student's mathematics portfolio than between the parts of a writing portfolio. Consequently, the reliability of mathematics portfolio scores was no better than that of the writing portfolio scores.[1]

Chapter Overview

This chapter addresses many of the same questions as the preceding one, and it largely follows the same organization. The next portion describes the procedures used to gather information about the degree of agreement between raters. We then discuss the rates of agreement between raters on individual pieces and dimensions. Subsequent sections examine the relationships among dimensions and pieces, estimate the contribution of various factors to the total variation in scores, and discuss the impact of alternative strategies for improving reliability. Issues discussed in the previous chapter that are not germane to this one include possible bias in teachers' ratings, systematic differences in scores among pieces, and comparisons to the reliability of Uniform Test scores.[2]

---

[1] One measure—simple percentage agreement between raters—appears better for mathematics than for writing, but we explain below why that is probably misleading.

[2] Classroom teachers were encouraged to score their own students' portfolios, but those scores were not used in the state reporting system evaluated here. In mathematics, all scored pieces were considered best pieces, and pieces were identified only by arbitrary position in the portfolio, so there is no reason to expect differences in average scores among them. And the Uniform Test in mathematics, unlike that in writing, was multiple choice, so its "rater reliability" was perfect.

Procedures

Mathematics portfolios were scored somewhat differently than writing portfolios. There was no distinction between "best" and "rest" in mathematics. Rather, each mathematics portfolio included 5 to 7 best pieces, and each of these pieces was graded on a 4-point scale on each of seven dimensions. Three dimensions were classified as pertaining to communication (language of mathematics, mathematical representations, and presentation) and were labeled C1 through C3. Four dimensions were classified as aspects of problem solving and were labeled PS1 through PS4 (see Appendix C for details). Thus, a rater assigned 35 to 49 scores to a portfolio (depending on the number of pieces) before grading the next portfolio.

The analyses described below were conducted on 803 Grade 4 portfolios and 344 Grade 8 portfolios that had at least 5 pieces scored twice. The second rater did not know the scores assigned by the first rater. Unlike writing, neither rater was the student's own classroom teacher. The 99 raters of Grade 4 portfolios and the 46 raters of Grade 8 portfolios participated in a training session before they began assigning final scores.

In analyzing the reliability of mathematics scores, we used two composite scores that were not used by the state for its reporting. First, we computed the average score for the portfolio on each dimension, across all pieces with valid scores. Second, we computed total score for the portfolio by summing these average scores across the seven dimensions.[3] These scores were computed separately by rater.

Agreement Between Raters

Because neither of the raters of mathematics portfolios was the student's own teacher, there was no opportunity to examine possible biases on teachers' evaluations of their own students. Agreement between raters was therefore only a question of the degree to which they were consistent with each other in deciding which pieces warranted high or low scores.

---

[3] The state used a complex algorithm, rather than a simple averaging, to obtain a composite score on each dimension. In this chapter, we focus on averages because they are simpler and more reliable. In Chapter 5, however, we use the state's composite scores, because the focus there is quality of aggregate scores that would be based on those composites.

The degree of agreement between raters was more variable in mathematics than in writing, but overall, it was quite low on both types of portfolios. This section presents the data on agreement rates on individual pieces, dimensions, and total mathematics portfolio scores.

Percent Agreement on Pieces

The degree of agreement between two raters on a single piece was generally higher in mathematics than on a single part in writing, but this difference is misleading because the mathematics agreement rates were still not much better than what would be expected by chance alone. That is, the rates of agreement were not much better than one would expect if each rater's score on one piece had been compared to a randomly selected piece that would usually be from a different student. As noted in Chapter 3, the rate of agreement expected by chance increases as scores become increasingly concentrated at one or two points on the scale, and mathematics scores tended to be more highly concentrated than writing scores.

Table 23 illustrates a typical pattern by showing the scores assigned by different raters to pieces on dimension PS2 ("How: Procedures") at Grade 4. The rows correspond to the scores assigned by the first rater and the columns to the scores assigned by the second rater. The entries in the table are the percentage of all twice-scored portfolios receiving each of the 16 possible combinations of scores. For example, 4.5% of the pieces that were read twice received a score of 1 from both raters. Overall, the two raters agreed on the score that should be assigned to 55.5% of the pieces (4.5 + 7.1 + 43.5 + 0.4 = 55.5). The 55.5% figure indicates raters agreed with each other slightly more than half the time on the score that should be assigned.

However, because scores were highly concentrated, the observed 55.5% rate is not much better than one would expect to occur by chance. Specifically, the likelihood that both raters would assign the same score to a piece by chance is the product of the rates at which they each assigned that score. In Table 23, for example, the likelihood that a piece would receive a score of 1 from both raters by chance is .124 x .119 = .015 or 1.5%. Similarly, the likelihood that both would assign a score of 2 is .233 x .219 = .049 or 4.9%. The corresponding calculation is made for the other two score levels, and the sum of the four products (1.5 + 4.9 + 38.4 + 0.1 = 44.9) is the overall rate of agreement that is likely to arise by

Table 23

Percentage of Grade 4 Mathematics Pieces Receiving Each Combination of Scores on the PS2 Dimension

		Score assigned by second rater				
		1	2	3	4	Total
Score	1	4.5	3.5	4.3	0.1	12.4
assigned	2	3.4	7.1	11.1	0.3	21.9
by first	3	3.9	12.5	43.5	2.6	62.5
rater	4	0.1	0.2	2.5	0.4	3.2
Total		11.9	23.3	61.4	3.4	100.0%

chance. In Table 23, the chance rate (44.9%) is only 10.6 percentage points less than the observed rate of 55.5%.

The rate of agreement between two raters on a piece varied across dimensions (Table 24). They ranged from 41% ("Why: Decisions" at Grade 4) to 86% ("What: Outcomes" at Grade 8). However, the higher rates of agreement typically reflected more substantial concentration of scores rather than evidence that raters could differentiate reliably among pieces of different quality; that is, even the higher rates of agreement were not much better than expected by chance.

Table 24

Mean Actual and Expected by Chance Agreement Rates on Mathematics Piece Scores, by Grade and Dimension

	Grade 4		Grade 8	
Dimension	Actual	Chance	Actual	Chance
C1-Language of Math	54	42	50	39
C1-Math Representations	48	33	52	33
C3-Presentation	45	33	45	30
PS1-Understanding of Task	66	57	65	55
PS2-How: Procedures	55	45	59	45
PS3-Why: Decisions	41	31	42	31
PS4-What: Outcomes	80	75	86	81
Average	56	45	57	45

Note: **Agreement rate = percentage of pieces that received the same score from both raters on a dimension.**

The most extreme concentration of scores was on dimension PS4 ("What: Outcomes of activities"). Table 25 shows that 86% of the Grade 8 pieces received the same score from both raters on this dimension. However, this seemingly high rate of agreement occurred because nearly all the eighth-grade pieces were given a score of 1 on this dimension. Specifically, 89% of the students were given a score of 1 by the first rater, 90% were given a score of 1 by the second rater, and 83% received a score of 1 from both raters. A few other criteria also had high degrees of score concentration (but not as high as the one in Table 25). For example, about 71% of the Grade 8 students were given a rating of 3 on the "Understanding of task" dimension.

A high rate of agreement when scores are highly concentrated may be due to reliable scoring or chance. In the case of PS4 "What: Outcomes," chance would have produced an overall agreement rate of 81% , only trivially lower than the 86% actual agreement shown in Table 25.[4] This is analogous to throwing darts at a target: If the bull's-eye is made large enough that almost all darts hit it, a high proportion of bull's-eyes no longer indicates which players can throw darts accurately. In the case of the PS4 dimension, a score of 1 is the bull's-eye, and the fact that almost all pieces get a 1 does not indicate whether raters can reliably differentiate among pieces deserving scores of 1, 2, 3, or 4.

To sum up, two raters often disagreed with each other on the score that should be assigned to a piece on a given dimension. When they did agree with

Table 25

Percentage of Grade 8 Mathematics Pieces Receiving Each Combination of Scores on the PS4 Dimension

		Score assigned by second rater				
		1	2	3	4	Total
Score	1	83.2	4.9	0.7	0.1	88.9
assigned	2	6.2	2.6	0.7	0.0	9.5
by first	3	0.8	0.3	0.3	0.0	1.4
rater	4	0.1	0.1	0.0	0.0	0.2
Total		90.3	7.9	1.7	0.1	100.0%

[4] 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 .90%.

each other frequently, it was because of extreme concentration of scores at one or two score levels. On all seven dimensions, the degree of agreement between raters was only slightly greater than what might have occurred by chance alone. In short, the data provide no evidence that raters could distinguish reliably between pieces that differed in quality on any of the seven dimensions.

Percent Agreement on Dimensions

Up to this point, we have discussed the degree of agreement between two raters on a single piece in a mathematics portfolio. We now examine agreement rates between raters on a single dimension for the portfolio as a whole. Were the raters consistent in saying that on a particular dimension, a given portfolio was one of the best, one of the worst, or somewhere in the middle relative to all of the other portfolios they graded?

To investigate this issue, we computed each portfolio's mean score (over its 5 to 7 pieces) on a dimension. Next, we rank ordered the means assigned by the first rater on this dimension from the highest to lowest, and then divided this distribution into four equal parts—the highest 25% (i.e., the top quartile), the next highest quartile, etc. We then repeated this process for the second rater.

These calculations, which were done separately for each dimension, allowed us to examine the degree of agreement between raters on a dimension. If raters agreed perfectly with each other regarding the grade that should be assigned to a student on a dimension, then every student's quartile on the first reading on that dimension would be the same as that student's quartile on the second reading—100% agreement. However, if there was no consistency between raters on this dimension, then by chance, 25% of the students would still be in the same quartile on both readings.

There was very little agreement between raters on a dimension even though each dimension score was based on 5 to 7 pieces. Table 26 illustrates the typical pattern. The first column shows that at Grade 4 on the "C2 — Mathematical Representation" dimension, only 36% of the students were in the same quartile on both readings (i.e., only 11 percentage points better than chance)—64% changed one or more quartiles. In fact, 5% of the portfolios changed three quartiles; that is, they went from the very bottom quartile on one reading to very top quartile on the other reading.

Table 26

Percentage of Grade 4 Students Whose Relative Standing Changed 0, 1, 2, or 3 Quartiles When a Different Rater Graded the Portfolio: Results for a Typical Dimension and Total Score

Amount of change	Typical dimension[a]	Total score across all dimensions
No change	36	45
1 Quartile	37	43
2 Quartiles	22	10
3 Quartiles	5	2

[a] Mathematical representation.

Percent Agreement on Total Scores

The last column in Table 26 shows that even when the analysis is based on total scores across all pieces and dimensions, 55% of the students changed one or more quartiles between readings. Taken together, Tables 23–26 show that the degree of agreement between raters on pieces, dimensions, and total scores is only slightly better than what is likely to occur by chance alone.

Correlation Coefficients

The correlation coefficient is another measure of the extent to which raters agreed with each other in their assessment of the relative quality of students' work. According to this index, raters often disagreed with each other as to which students did better than others on a piece, on a dimension, or even the whole portfolio.

Unlike simple agreement rates, correlation coefficients are not inflated by the raters assigning the same score to almost all the pieces. For example, because almost all the pieces received a score of 1 on the PS4 "What: Outcomes" dimension in Grade 8, the raters agreed with each other 86% of the time on the score that should be assigned. However, the correlation coefficient between two raters on a piece on this dimension was only .30. This is actually slightly below the typical correlation between raters at the piece level on the other dimensions (see first two columns of Table 27).

Combining information across pieces increased the correlations between raters, but only marginally. As noted earlier, each portfolio's mean score on each

Table 27

Correlations Between Raters on Pieces and Dimensions

Dimension	Piece level		Dimension level	
	Grade 4	Grade 8	Grade 4	Grade 8
C1-Language of Math	.34	.32	.30	.28
C1-Math Representations	.41	.47	.36	.34
C3-Presentation	.39	.45	.51	.53
PS1-Understanding of Task	.30	.32	.42	.38
PS2-How: Procedures	.33	.36	.48	.38
PS3-Why: Decisions	.35	.37	.48	.37
PS4-What: Outcomes	.30	.31	.43	.39
Mean	.34	.37	.42	.38

Note. **The piece-level columns show the correlations between the first and second raters in the scores they assigned to individual pieces. The dimension-level columns show the correlations between raters in the mean scores they assigned to a portfolio on a dimension where a dimension score is the mean over 5 to 7 pieces.**

dimension (across its 5 to 7 pieces) was computed separately by rater. The correlations between the means assigned by the first and second raters on a dimension ranged from .28 to .51 (see last two columns of Table 27). Averaged across all dimensions, the means of these dimension-level correlations between raters were .42 in the fourth grade and .38 in the eighth grade, only slightly higher than the corresponding correlations at the level of individual mathematics pieces (.34 and .37). These dimension-level correlations were quite similar to those found for writing: .39 in the fourth grade and .49 in the eighth grade.

Summing scores across all dimensions to yield a single total score per portfolio improved reliability modestly relative to the dimension-level scores, but even these total-score correlations were only moderate. At Grade 4, the total score assigned to a portfolio (across all dimensions) by one rater correlated .60 with the total score assigned by the other rater. At Grade 8, the correlation was .53. These values are similar to the corresponding correlation coefficients on the writing portfolios (.49 and .60, respectively).

Relationships Among Scores on Different Pieces

All of the scored pieces in the mathematics portfolios were considered best pieces. Moreover, there was nothing to distinguish one piece from another beyond their arbitrary positions in the portfolio. Thus, this section considers only the correlations in scores across pieces.

Scores for the various pieces of a mathematics portfolio typically had lower correlations with each other than did the scores on the two parts of the writing portfolios. The simple correlations between pieces were extremely low when they were scored by different raters. Across dimensions, the correlations were .10 or less, and within a single dimension, they were .13 or lower. (These are the raw correlations in Table 28.) Even when scores were assigned by the same rater on the same dimension, the correlations between pieces of the math portfolio were low, averaging only .27. In contrast, the corresponding raw correlations in writing were all two to three times as large (Table 14).

The much lower correlations in mathematics compared to writing cannot be attributed to differences in rater reliability. When disattenuated for rater reliability, the mean correlation between raters on a mathematics piece in Grade 4 was .39 for the same dimension and .29 across different dimensions. Similar correlations were obtained at Grade 8. The corresponding correlations in writing were .77 and .73 in the fourth grade and .89 and .81 in the eighth grade (Table 14).

Table 28

Mean Observed and Disattenuated Correlations Between Mathematics Pieces When Scores Are Assigned by the Same or Different Raters

	Grade 4		Grade 8	
	Raw	Corrected	Raw	Corrected
Different raters				
Same dimension	.13	.39	.12	.33
Different dimensions	.10	.29	.09	.27
Same rater				
Same dimension	.27	—	.26	—
Different dimensions	.19	—	.21	—

Note: **Disattenuated correlations cannot be computed for the same rater.**

It is not clear why the correlations in mathematics were much lower than they were in writing, but several related possibilities present themselves. Mathematics pieces within a portfolio may have been, on average, more dissimilar to each other than were the pieces comprising writing portfolios. Some mathematics problems, in contrast to essays, may have relatively clear correct answers (in the Vermont program, "clear solutions" or "clear ways of presentation" might be more appropriate). Although phrased in generic language, the scoring rubrics may have been more clearly applicable to some pieces within a portfolio than to other pieces. The available data were not sufficient, however, to evaluate these or other possible explanations.

Relationships Between Scores on Different Dimensions

Mathematics scores for a given piece showed low correlations across dimensions. Regardless of grade level, the average correlation between dimensions for a single piece was roughly .20 when different raters assigned the scores, and .35 when the same rater assigned the scores (Table 29).

Even though correlations between dimensions at the individual piece level were typically low, the data suggest that many of the dimensions may not be independent at the level of the total portfolio. The most reliable score we could obtain on a dimension for a mathematics portfolio was produced by taking the mean of all of the scores assigned to it—that is, the scores assigned to all pieces by both raters. These correlations averaged a bit over .50 (Table 30). When these correlations are disattenuated for rater unreliability, all but one of the dimensions (PS4 — "What: Outcomes of activities") were highly correlated with each other, with many correlations above .90 (Table 31). These correlations suggest that with the exception of PS4, there was not much underlying difference among the dimensions in the 1991-92 total portfolio scores. However,

Table 29

Mean Correlations Between Dimensions When Scores Are Assigned by the Same Versus Different Raters

	Mathematics		Writing	
	Grade 4	Grade 8	Grade 4	Grade 8
Different rater	.20	.20	.29	.37
Same rater	.35	.36	.57	.61

Table 30

Interdimension Correlations Math, Unadjusted

	C1	C2	C3	PS1	PS2	PS3	PS4
C1		.41	.63	.58	.58	.55	.43
C2	.36		.47	.47	.53	.46	.32
C3	.57	.55		.72	.75	.83	.39
PS1	.45	.51	.73		.81	.72	.39
PS2	.50	.56	.75	.79		.78	.35
PS3	.57	.50	.81	.66	.77		.29
PS4	.36	.34	.36	.36	.35	.33	

Note. **Data for Grades 4 and 8 appear below and above the main diagonal, respectively.**

Table 31

Interdimension Correlations Math, Adjusted

	C1	C2	C3	PS1	PS2	PS3	PS4
C1		.87	*	*	*	*	.86
C2	.73		*	*	1.00	.89	.59
C3	*	.93		*	*	*	.62
PS1	.86	.92	*		*	*	.71
PS2	.92	.95	*	*		*	.63
PS3	*	.88	*	*	*		.53
PS4	.68	.59	.60	.56	.54	.33	

Note. **Data for Grades 4 and 8 appear below and above the main diagonal, respectively.**

*** = Adjusted estimate of correlation is greater than 1.00.**

this extreme a correction for disattenuation is risky, and these disattenuated correlations are only uncertain estimates of what would have been found if raters scored reliably.

Sources of Variation: A Generalizability Analysis

As with writing portfolios, we used generalizability analysis to explore how much of the total variation in mathematics portfolio scores was attributable to various sources: students, raters, pieces, interactions among these factors, and

"noise" (residual error variance). With one very important exception, the results in mathematics paralleled those in writing. Specifically, just as in writing:

- **About half of the variance in mathematics scores on a dimension was due to unsystematic inconsistencies between raters: The combination of the Student x Rater interaction and residual error variance accounted for 53% of the total variance at Grade 4 and 52% at Grade 8 (Table 32).**
- **There was relatively little systematic difference in leniency among raters (see the rater effect in Table 32).**

However, unlike writing, consistent differences among students in total portfolio scores (the main effect for students) accounted for only 15% of the variance in the fourth grade and 13% in the eighth grade. In contrast, the main effect of students in writing accounted for roughly 30% of the variance. Moreover, about one-fourth of the variance in mathematics scores was due to a Student x Piece interaction. This indicates that students received higher scores from both raters on some pieces than on other pieces in their portfolios; that is, the students themselves were not consistent in their performance level across pieces. The relatively large Student x Piece interaction corresponds to the finding reported earlier that the correlations between pieces in mathematics portfolios were smaller than the correlations between parts in writing portfolios. This means that scores on a single piece are a less trustworthy measure of student proficiency in mathematics than in writing. Regardless of whether the variability in performance across pieces is good news or bad in other respects—

Table 32

Percentage of Variance in Mathematics Portfolio Scores on a Typical Dimension That Was Attributable to Various Factors

Source	Grade 4	Grade 8
Students	15	13
Students x Pieces	23	27
Raters	10	7
Raters x Students	5	8
Residual	48	44
Total	100%	100%

Note. **Values may not sum to 100% because of rounding.**

and it could be either or both—it does indicate that a reliable measure of performance is likely to require more pieces in mathematics than in writing.

Strategies for Improving Score Reliability

The sections above focused primarily on the reliability of scores assigned to specific pieces and dimensions. In this section, we examine the effects of combining scores across pieces and dimensions. We also estimate the effects of changing the number of pieces in the portfolio and the number of times each piece is graded by a different rater.

Effects of Averaging Scores Across Pieces and Dimensions

Combining scores across pieces and dimensions in mathematics produced only modest gains in reliability (Table 33). The degree of agreement between raters was still fairly low even when scores were summed over all pieces and dimensions. The increases in reliability from each type of combining are quite similar to those in writing (Table 18), as is the maximum reliability obtained by combining. In mathematics, the maximum reliability coefficients were only .60 and .53 in the fourth and eighth grades, respectively. The corresponding correlations in writing were .49 and .60.

Increasing the Reliability of Total Scores by Adding Pieces or Raters

Just as in writing, increasing the number of independent raters who evaluate a portfolio from 1 to 2 had a noticeable effect on the reliability of total scores regardless of the number of pieces within the portfolio that are evaluated.

Table 33

Average Correlation Between Raters With Different Types of Combining, Mathematics and Writing

	Mathematics		Writing	
	Grade 4	Grade 8	Grade 4	Grade 8
No combining—one dimension on one piece[a]	.34	.37	.34	.43
One dimension—mean over all pieces	.42	.38	.39	.49
One piece—mean over all dimensions	.48	.50	.45	.56
Total—mean over all dimensions and pieces	.60	.53	.49	.60

[a] In writing, piece refers to parts, see Chapter 3.

But again, there is not much to be gained by having more than 2 raters or more than 5 pieces evaluated. However, even under these conditions, the reliability of a student's total score across all pieces and dimensions is still only .60. Figure 2 shows the trade-offs among the number of independent raters per portfolio and number of pieces evaluated for one of the math dimensions. (Similar figures for all of the math dimensions are included in Appendix C.)

Conclusions

The results presented in this chapter show that the scores assigned to mathematics portfolios were just as unreliable as the scores assigned to writing portfolios, and the low to modest agreement between raters was again the major source of the problem. On a given piece, the degree of agreement between two raters was only slightly better than what would occur by chance alone. The mean score assigned to a portfolio on a dimension by the first rater had only a low correlation with the mean assigned by the second rater. The same was true for the total scores assigned by each rater (i.e., across all dimensions and pieces). Overall, the mathematics raters were no more consistent with each other than were the writing raters, and neither group of raters provided adequately reliable scores.

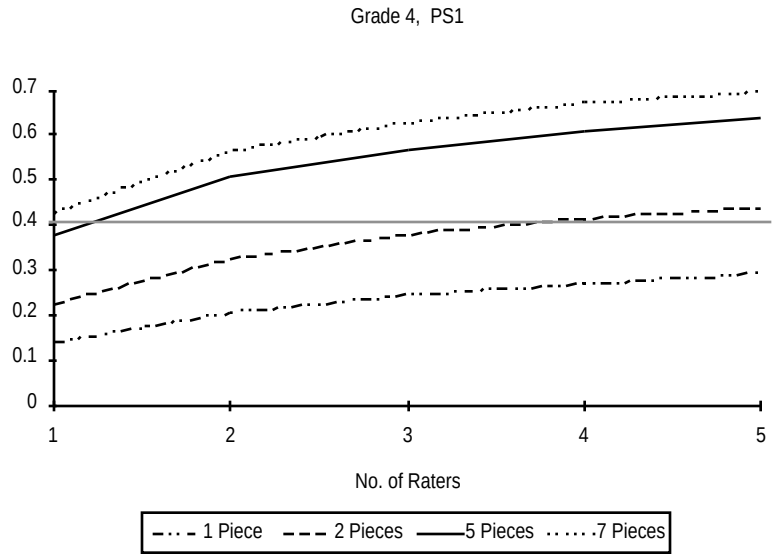


Figure 2. **Effects on reliability of increasing the number of raters or pieces, Grade 4, PS1.**

As in writing, there was no apparent value to scoring the portfolios on separate criteria (dimensions) because after adjusting for the low reliability of the raters, the correlations among the dimensions were near perfect. The sole exception to this trend was that the "PS4 What: Outcomes" dimension appeared to measure a somewhat different aspect of a student's work than did the other dimensions.

In writing, students who received a relatively high score on the best part were likely to receive a high score on the rest. That did not happen in mathematics. Instead, there was a relatively low correlation among the separate pieces (i.e., there was a large Student x Piece interaction). This is the major reason why the total score on a mathematics portfolio was not more reliable than the total score on a writing portfolio (even though many more scores were assigned to each mathematics portfolio).

Finally, increasing the number of raters per portfolio from 1 to 2 will increase reliability by a noticeable amount, but using more than 2 raters will not help much. Adding pieces helps too, but there are substantially diminishing returns in reliability by using more than a half-dozen pieces. More importantly, there is no operationally feasible combination of raters and pieces that will provide an acceptable level of total score reliability for a mathematics (or a writing) portfolio.

CHAPTER 5. THE QUALITY OF AGGREGATE SCORES

The Vermont portfolio assessment program is designed to provide aggregate information, including statewide data and information about smaller aggregates, such as supervisory unions, districts, and schools.[1] The quality of such aggregate data depends in part on the reliability of individual scores; the low rater reliability described in the previous two chapters will decrease the reliability of aggregate statistics. Depending on the specific aggregate statistics in question, however, a variety of additional factors come into play as well, including:

1. ***Sampling error.*** **In some cases (e.g., eighth-grade mathematics), scores were available only for a sample of the state's students.[2] 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 in estimates for the entire state but is primarily a problem for statistics from smaller groups, such as schools or supervisory unions.[3]**

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. Simple estimates of error assume that all observations (in this instance, students) are independent draws from the population,**

---

[1] Teachers may use portfolio scores for evaluating individual students, but no reporting or use of individual students' scores outside of the school have every been planned.

[2] In the 1992-93 school year, all reporting will be based on samples of approximately 1,750 portfolios per grade in each subject.

[3] 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. Therefore, even if all students in a school have their portfolios scored, the resulting data do not provide an error-free estimate of the level of performance for that school. Rather, the data provide an estimate of the performance *of that cohort* in that school. In technical terms, this is the reason that we did not apply finite-sample corrections in estimating the error of aggregate scores.

so those estimates of error should be increased to take clustering into account.

3. *Biased estimates of proportions.* As explained below, unreliability of scoring will generally result in too many students obtaining extreme scores. This undermines a number of aggregate statistics that might otherwise be useful.

In the following sections, we explore the impact of these factors on the quality of both statewide and school-level scores.

Statewide Scores

Despite the unreliability with which the work of individual students was scored, some statewide statistics, such as average scores on a dimension, were sufficiently reliable to report because of the relatively large numbers of portfolios rated statewide. However, other potentially useful statistics, such as the proportion of students at a given score point, were rendered unusable. Moreover, in judging the reliability of aggregate statistics, as in evaluating the reliability of individual portfolio scores, we considered only one component of reliability: the consistency of ratings. In this chapter, we make no assertions about the reliability of aggregate scores in any broader sense, such as the consistency of scores across instances of measurement. For example, we were not able to assess the degree to which aggregate scores (such as the ranking of schools) would have been affected by substitution of different tasks in students' portfolios.

Average Scores

Statewide average scores were reasonably precise, as illustrated by fourth-grade mathematics composite scores (Table 34). The first column provides the average score on each criterion, and the second column indicates how far the margin of error extends *in each direction.* These margins of error are twice the standard error.[4] This is referred to as the margin of error. For example, the average score on "Language of mathematics" was 1.7 out of a possible 4, and the

[4] The range extending from the average minus the margin of error to the average plus the margin of error is approximately a 95% confidence band. In this case, and elsewhere when appropriate, confidence bands were estimated by a school-level jackknife procedure to reflect clustering.

Table 34

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

margin of error extended .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 35).

Despite the small margins of error, however, statewide averages for 1991-92 still had serious limitations. The most important is that some districts and schools opted out of the program. In the case of fourth-grade writing, 22 of 246 schools did not contribute portfolios. This represents a relatively small percentage (about 9%), but it is not random; for example, the largest district in the state (Burlington) withdrew from the program before scoring. Accordingly,

Table 35

Average Eighth-Grade Writing "Rest" 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

statewide averages must be interpreted as representing only participating schools and districts. Probably less important, portfolios were sampled (on both a planned and an *ad hoc* basis) to compensate for a shortage of raters. We have not estimated the likely effects of the nonrepresentativeness that might have resulted from factors such as these.[5]

Proportions of Students at Each Score Point

The Vermont State Department of Education has been less interested in reporting averages than in reporting the proportion of students reaching each of the 4 scale points on each scoring dimension. There are several reasons for this preference. In theory, the proportion of students reaching each score provides more useful diagnostic information than does a simple average. Some also believe that avoiding averages may help lessen the "horse race" nature of comparisons among schools or districts.

Unfortunately, the proportion of students reaching each score is a more problematic statistic under the best of circumstances, and the low rater reliability documented in the previous chapters made it unusable in 1991-92. There are two problems: large margins of error and bias.

Margins of error. In general, reporting of proportions will be more difficult than reporting of averages. The margins of error will often be larger than those of averages. Moreover, accurate estimates of the margins of errors for proportions will be difficult to obtain. Sampling error and clustering can be addressed straightforwardly in the case of proportions, as in the case of averages. However, both measurement error at the level of individual students and bias in estimated proportions (discussed below) make it difficult to estimate the margin of error in proportions.

At the statewide level, proportions would have a sizable margin of error even if individual portfolios were scored with perfect reliability (Table 36).[6] In the case of Presentation, for example, the margin of error would be +/- 4 percentage points. Consequently, it would be more reasonable to present the

---

[5] In contrast, in 1992-93, a planned random sample of portfolios was scored in both grades and subjects.

[6] These estimates reflect all 1,957 and 1,955 portfolios for which we had valid scores on each of these dimensions, but the margins of error are considerably larger than the conventional formula would suggest because of the clustering of scores at the school level.

Table 36

Observed Proportions of Students at Each Score Point and Margins of Error, Fourth-Grade Mathematics, Understanding and Presentation (Assuming Perfect Rater Agreement)

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

proportion of students scoring 3 as "43% to 51%" or "roughly half," rather than "47%."

At the school level, the margin of error for proportions would be very large because of the smaller numbers of students. To illustrate this, Table 37 provides the margins of error for a proportion of 20%, assuming perfect rater reliability, for groups of different sizes. Thus, if a school includes 24 fourth-grade students whose portfolios are scored (the median fourth-grade enrollment among schools in Vermont in 1991-92) and 20% of them receive a certain score, the margin of error around the estimate of 20% extends from 4% to 36%.[7]

Table 37

Margins of Error for Proportions of 20%, by Number of Students (Assuming Perfect Rater Agreement)

Number of Students	+/- Error (Percentage Points)
15	21
24[a]	16
30	15
45	12
60	10
100	8

[a] **Median fourth-grade enrollment for Vermont schools in 1991-92.**

[7] **These are simple random sampling estimates of twice the standard error of a proportion. In part because of the small size of many Vermont schools, we assumed no clustering within schools.**

Worse yet, the margin of error for differences between schools is larger—often by a factor of nearly 1.5—than the margin of error for estimating the proportions within a single school. (This is explained further in the following section on school-level scores.) Thus, comparisons of proportions among small schools (or other small groups) based on a single year of scores is simply not practical.

Bias. When scores are unreliable, they tend to spread out more than they would if scoring were reliable.[8] This leads to biased estimates of the proportion of students achieving each score. Too many students receive very high and low scores (in this case, 1 and 4), and too few receive scores near the middle (2 and 3).

In the 1991-92 portfolio program, the unreliability of ratings of individual portfolios was sufficient to cause serious bias in the proportions reaching each score, even at the statewide level. We estimated "true" proportions for two criteria in fourth-grade mathematics, "Understanding" and "Presentation."[9] 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 observed 2% and 1% (Table 38). The more substantial bias, however, was in the scores of 2. 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 recommended against reporting either observed or estimated true proportions for 1992, and the Vermont State Department of Education followed that recommendation.

---

[8] This assumes that true scores become less frequent toward the extremes of the scale.

[9] These estimates are based on the assumption that the unobserved latent scores have a bivariate normal distribution. Measurement error was estimated by method of moments—i.e., by computing the interrater correlation for latent scores necessary to produce the observed correlation for the interrater data. Monte Carlo methods were used to estimate the relationship between these two correlations (40,000 draws per point). This procedure introduced some error into the estimates but that error is small compared with the sampling error in estimating the interrater correlations for observed scores.

Table 38

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

School-Level Scores

Scores (such as averages) for small groups are unreliable even when raters show perfect agreement on scores for students simply because of the small number of students contributing to each aggregate score. In the case of the Vermont portfolio program, this general problem was exacerbated by unreliable scoring for individual students, but the particularly small size of many Vermont schools would result in large margins of errors for average scores even if ratings were completely reliable.

The following statistics, drawn from enrollment data for 1991-92 provided by the state Department of Education, illustrate the severity of the small-school problem in Vermont:

Number of schools with fourth-grade students: 246.

Number of schools from which we had fourth-grade writing portfolio scores: 224.

Minimum number of fourth graders enrolled: 1.

Median number of fourth graders enrolled: 24.

Mean number of fourth graders enrolled: 33.

Maximum number of fourth graders enrolled: 170.

As noted above, the enrollments in many of these schools are small enough that proportions of students reaching each score are too unreliable to use. Even simple averages, however, are unreliable in the smaller of these schools, and they would not be much improved even if rater reliability became perfect. To

illustrate this, Table 39 presents confidence intervals around means from three actual (but unnamed) schools chosen on the basis of enrollments. The table lists the number of portfolios scored in writing, the mean "Purpose" score, and three confidence intervals. The first confidence interval reflects the actual reliability of scoring reported above. The second confidence interval is what would obtain if the reliability coefficient (Spearman's rho) were increased to .60, which is probably a reasonable target for the next year or two. The final column shows the confidence interval for perfectly reliable scoring; this is unattainable for subjective scoring but provides a best-case view of the reliability of school means.[10] Thus, for example, in School One the 95% confidence interval extends 0.4 to either side of the mean score of 3.3, given the observed reliability of rating—that is, from 2.9 to 3.7. Even perfectly reliable scoring (r=1.0) would only shrink the confidence interval modestly, from 0.4 to 0.3.

These are the confidence intervals one would use if one were drawing inferences only about the average score in a single school. If the issue is differences between scores, one would need to use a larger ("simultaneous") confidence interval. These larger confidence intervals for the same three schools are shown in Table 40.[11]

Thus, for example, in the case of School One, the observed mean is 3.3, but the mean from a second school of similar size would have to reach 3.9 for us to have confidence that it is really better. Even with a reliability of .60, which would be a large improvement from this past year, a second score would have

Table 39

Confidence Bands for Purpose, Grade 4 Best Piece

School	Number of students	Mean score	+/- error, observed	+/- error, r = .60	+/- error, r = 1.0
One	14	3.3	0.4	0.3	0.3
Two	31	2.7	0.2	0.2	0.1
Three	53	3.1	0.2	0.1	0.1

[10] Note that some changes in the confidence intervals as reliability is increased do not appear in Tables 39 and 40 because of rounding.

[11] These assume that the second school in each comparison is identical in size and in the internal distribution of scores.

Table 40

Simultaneous Confidence Bands for Purpose, Grade 4 Best Piece

School	Number of students	Mean score	+/- error, observed	+/- error, $r = .60$	+/- error, $r = 1.0$
One	14	3.3	0.6	0.5	0.4
Two	31	2.7	0.3	0.3	0.2
Three	53	3.1	0.3	0.2	0.1

to be at least 3.8 for us to be confident that it is really higher. With perfect reliability of ratings, the second score would have to be at least 3.7. Recall that more than a fourth of Vermont's schools are at least as small as School One. School Two is considerably larger—recall that 60% of Vermont's fourth grades are as small as School Two's—but even with moderate rater reliability, we could be confident that a second mean is different from School Two's average only if it was at least 3.0, compared to School Two's mean of 2.7.

The differences in averages that are needed to be confident that two schools really differ at all are quite large relative to the observed differences among schools. In 1991-92, we found the following distribution of school means on Purpose for fourth-grade best pieces:

1.8 Minimum

2.8 25th Percentile

3.0 Median, Mean

3.2 75th Percentile

4.0 Maximum

Thus School Two in Tables 39 and 40, with a mean of 2.7, scored below the 25th percentile last year. To be precise, it scored at the 20th percentile of the 224 schools for which we had fourth-grade writing portfolio scores. Yet only schools that are average or above average can be considered with any confidence to have scored higher than School Two.[12]

---

[12] This is an oversimplification, because the comparisons between School Two and other schools depend on the size and distributions within the other schools as well.

Conclusions

These results show that the unreliability with which the portfolios of individual students were rated seriously limits the uses to which the 1991-92 portfolio data can be put. Statewide averages had reasonably small margins of error, but statewide estimates of the proportion of students reaching each score point were both biased and unreliable and cannot be used. At the level of individual schools, even averages had such large margins of error that only very large differences between schools were reliable.

It is important to realize, however, that attaining higher levels of rater agreement will not solve all of these problems. Because of the small size of many Vermont schools, many comparisons of average scores among them—if based only on a single year's scores on a single assessment—would be unreliable even if individual portfolios were rated with perfect agreement. Comparisons of averages among the smaller districts and supervisory unions will be similarly untrustworthy. Moreover, because margins of error are poorly understood, publication of those averages could lead people to make unwarranted conclusions—for example, to conclude that chance differences in scores are real. Finally, because the reliability of portfolio scoring will likely improve only gradually, estimates of the proportions of students reaching each score point will probably remain problematic.

CHAPTER 6. VALIDITY

Rater reliability is a necessary but insufficient basis for judging an assessment to be successful as a measurement tool. The ultimate question is the validity of the assessment: evidence that the data it yields support the inferences about student performance that people base upon it. Unreliable scoring undermines validity, but reliable scoring does not guarantee it.

It has become common to include under the rubric of "validity" evidence of other effects, such as effects on instruction or on the equity of educational services. These effects are often labeled "consequential validity" or "systemic validity." In this report, however, we use "validity" in its traditional and narrower sense, that is, to refer to the quality of the data produced by the assessment. We do not mean to downplay the importance of other consequences of an assessment. On the contrary, such other effects are a primary rationale for the Vermont program and are a major focus of our evaluation. (For example, effects on the mathematics curriculum are described in Chapter 2.) It is simply clearer to use other terminology to refer to those effects and to reserve "validity" for a discussion of the meaningfulness and interpretability of the assessment's data.

In 1991-92, the question of the validity of portfolio scores was largely mooted by the very low rater reliability. However, we explored evidence of what the validity might have been if reliability had been higher. Even though estimates of how scores might have behaved in the absence of error are risky, particularly when the error is as large as it was in the Vermont portfolio system, the effort was instructive. It provided some initial hints about the validity of the Vermont portfolio scores. It also illustrated a number of impediments to effective validation, not only of the Vermont program specifically, but also of many performance assessment systems.

Criteria for Validating the Vermont Program

A wide variety of evidence can be adduced to test the validity of an assessment. In 1991-92, we focused on both evidence from the scores themselves and evidence from our investigation of the implementation of the program.

One criterion we examined is generalizability of performance: the consistency of students' performance over alternative measures of the same construct or achievement domain. The validity of inferences about students' performance in a given domain would be undermined by evidence that their performance does not generalize well across measures. For example, suppose that one had two alternative tests of algebra, judged to be roughly equivalent in their difficulty and their representativeness of the domain of algebra. If students' performance on one of the tests was inconsistent with performance on the other—that is, if their performance did not generalize across the two tests—one could not consider either one a valid basis for judging students' ability in algebra, because they would suggest different conclusions about which students had mastered the subject. Generalizability is a matter of reliability as well as validity and is sometimes called "score reliability."

We also looked at "convergent" and "divergent" evidence. These cumbersome terms refer to a fairly simple notion: Scores on a test should correlate more highly with measures of highly related constructs than with measures of less related constructs. For example, proficiency in calculus should be more highly related to proficiency in other aspects of mathematics—say, trigonometry—than to vocabulary. Therefore, if scores on a test of calculus failed to correlate more highly with trigonometry scores than with vocabulary scores, one would have good reason to doubt the validity of the calculus scores.

For comparisons of this sort, we used scores from the state's uniform assessments of mathematics and writing. The Uniform Test of mathematics was a matrix-sampled test that included both multiple-choice and open-ended items.[1] However, the open-ended items were not scored, and the scores used by the state were based on the 30 multiple-choice items administered to each student. In our study of validity, we similarly used only the multiple-choice items. The writing Uniform Test was a single essay written in response to a single prompt used throughout the state. The essays were scored by employees of the state's testing contractor rather than by Vermont teachers, but the raters used the same scoring criteria and rubrics that were used to score the writing portfolios. We also collected mid-year and final mathematics grades and all standardized test scores for students in a stratified random sample of Vermont schools.

---

[1] A matrix-sampled test is one in which each student receives only a sample of the total pool of test items.

Interview and questionnaire data also provided evidence pertaining to validity. For example, as noted in Chapter 2, we found evidence pertaining to variations in the implementation of the program that might bear on the validity of scores. We also obtained useful feedback from portfolio raters.

Barriers to Validation

The barriers to validation we encountered were numerous. Some were idiosyncratic aspects of the Vermont program or of the context in which it operates, but others were factors that are likely to affect a variety of performance assessment systems.

As noted above, the first obstacle to validation was the low level of rater reliability we found; an unreliable measure cannot be valid. Moreover, the reliability of ratings was so low that it clouds our estimates of what relationships among scores would have been if rating had been better. There are techniques for estimating what relationships would be found if rater reliability had been high—the disattenuation methods used in Chapters 3 and 4—but when error is as large as it was in the case of the portfolio scores, the resulting estimates are uncertain.

Our sample schools used a variety of different standardized tests and employed a wide variety of grading methods and standards. Accordingly, we planned to conduct our analysis of both grades and standardized test scores within schools and then to pool the results across schools. The very small enrollments in many Vermont schools would have made that difficult at best, and the low reliability of scores would have exacerbated the problem of small numbers. The decision at the end of the year to score only a sample of mathematics portfolios from each school, necessitated by an insufficient number of raters, made most within-school samples too small for the planned analyses.

Of more general importance than these concrete problems, however, was the insufficiently clear definition of the domain of mathematics that the portfolio assessment was supposed to tap and the lack of a clear notion of the relationships that should obtain between that domain and others, such as more traditionally assessed aspects of mathematics and aspects of verbal fluency. (Recall that several of the seven mathematics criteria pertain to communication.) This is a problem that will plague many performance assessment programs. One

rationale for performance assessments is that they will measure different aspects of competence than do traditional tests. Moreover, many are intended to bridge more than one traditional domain. (The Vermont mathematics portfolio program, with its emphasis on written communication, is a clear example.) Thus, up to a point, a moderate correlation between mathematics portfolio scores and scores on a multiple-choice test might be construed as better news than a very high correlation. A very high correlation might signify that the portfolios were not providing much information beyond that available from the multiple-choice test, while a moderate correlation might indicate that the portfolios were successfully tapping other aspects of proficiency in mathematics. A substantially lower correlation, however, might be damning; it might indicate that the portfolio scores are too heavily influenced by things that are not germane. (For example, many teachers have expressed concerns that the emphasis on written communication in the mathematics portfolio program, whatever its instructional benefits, might be undermining the validity of the scores for fourth-grade students with relatively weak proficiency in verbal expression and writing.)

Thus, there is as yet no firm basis for deciding what would constitute good or bad news. Clear evidence of validity will require more clarity about the patterns of relationships that portfolio scores should show (particularly outside of the area of writing), and it will likely also require an expanded range of measures that can be used for comparison.

Evidence From the Writing Portfolio Scores

As noted in Chapter 3, scores on the two parts of the writing portfolios were quite consistent, after removing the effects of rater unreliability. Indeed, after disattenuating to remove the effects of unreliable rating, the correlations between the two parts of the portfolio were higher than some other research would have predicted.

Some raters might construe that consistency between scores on the two parts as evidence of generalizability of performance. There are several reasons to be cautious, however. The portfolio program may have produced a limited and nonrepresentative sample of the domain of writing. Even if the program as a whole sampled reasonably well from the domain, it is quite possible that the portfolios of individual students did not. (Recall that in mathematics, we found evidence of marked between-school differences in task selection.) Moreover, the

extremely low level of rater reliability makes the disattenuation suspect. It is possible that if raters had scored reliably, the correlations shown by their scores would have been different from our disattenuated estimates.

For these reasons, we looked at the state's Uniform Test of writing for additional convergent evidence of validity. When we compared scores on the writing portfolio to scores on the writing Uniform Test, we obtained lower correlations than we found between the parts of the portfolio. The raw correlations were very low, as the low rater reliability of the portfolio scores preordained. In Grade 8, for example, the average correlation (across dimensions) between the best piece and the rest of the portfolio was .37; the average correlations between the Uniform Test and the best and rest were .31 and .28, respectively. After disattenuating for rater unreliability, however, the correlations between scores on the portfolio and the Uniform Test were moderate, ranging from .47 to .59, compared to .80 or higher for the correlation between the best piece and rest (Table 41). The rest scores showed a bit higher correlations with the Uniform Test than did the best piece. It is possible that the larger number of pieces of work entering into the rest scores made those scores a bit more robust.

These disattenuated correlations are reasonable, in the light of other research.[2] They suggest that although much of the variation in ratings is

Table 41

Disattenuated Correlations Between Writing Pieces, Portfolio and Uniform Test

	Portfolio: best piece vs. rest	Portfolio (rest) versus Uniform Test	Portfolio (best) versus Uniform Test
Grade 4	.80	.59	.47
Grade 8	.86	.58	.52

Note. **Averages across dimensions of correlations calculated within dimensions.**

[2] **The Uniform Test was a single prompt, so the correlations between the best piece and the Uniform Test are analogous to the correlations between single essays reported in other research (see Dunbar, Koretz, and Hoover, 1991).**

error, raters are to some degree recognizing differences in the quality of writing. However, only more reliable scoring will permit us to test whether this inference based on disattenuation is correct.

For divergent evidence, we also compared writing portfolio scores to scores on the mathematics Uniform Test. Because we considered only scores from multiple-choice items and ignored items that required writing, one would expect that the writing portfolio scores would show a lower correlation with scores on the math Uniform Test than with scores on the writing Uniform Test. This was not the case. In the fourth grade, writing portfolio scores showed nearly identical correlations with the two Uniform Tests (Table 42). In the eighth grade, correlations with the writing Uniform Test were higher than those with the mathematics test, but only marginally. These correlations, however, are ambiguous because of the limited scope of the writing uniform assessment. That is, the fact that the writing uniform assessment comprised only a single prompt would tend to depress the correlations between that test and the writing portfolios.

Evidence From the Mathematics Portfolio Scores

Correlations between scores on the mathematics portfolio and the mathematics Uniform Test are also difficult to interpret because of both the severity of the disattenuation for rater error and the lack of a clear expectation for the relationships that should obtain between the portfolio scoring dimensions and other aspects of performance in mathematics. Nonetheless, there is little in the correlations found in 1991-92 to generate confidence in the portfolio scores.

Table 42

Disattenuated Correlations Between Writing Portfolio Scores and Uniform Test Scores in Writing and Math

	Writing Uniform Test	Math Uniform Test
Grade 4		
Best	**.47**	**.50**
Rest	**.59**	**.61**
Grade 8		
Best	**.52**	**.43**
Rest	**.58**	**.52**

As noted in Chapter 4, scores on the mathematics pieces within a portfolio were less consistent with each other than were scores on the two parts of the writing portfolio. The average raw correlations between the pieces varied from one dimension to another, but all were very low; 11 of the 14 correlations were less than .20. Even after disattenuating for rater unreliability, the average correlations between pieces remained low: The overall average (across dimensions) was .38 in fourth grade (Table 43, left-hand column). Eighth-grade results were similar; the overall average (across dimensions) was .31.

Accordingly, we compared composite scores from the portfolios, rather than piece-level scores, to scores on the mathematics Uniform Test. Given the low correlations among piece-level scores, the composite score should be more reliable and should show higher correlations with other measures of performance in mathematics.

Even after disattenuation, the correlations between mathematics portfolio composite scores and the math Uniform Test were typically quite low, averaging about .32 (Table 43, right-hand column). These correlations showed a different pattern across dimensions than did the correlations among the portfolio pieces themselves (Table 43, left-hand column), but the average correlation is lower.

Table 43

Average Disattenuated Correlations Between Mathematics Portfolio Pieces and Math Uniform Test Scores, Grade 4

Dimension	Portfolio pieces with each other	Portfolio composite with Uniform Test
Language	0.21	0.42
Representations	0.15	0.31
Presentation	0.54	0.36
Understanding of task	0.39	0.41
How: Procedures	0.42	0.33
Why: Decisions	0.57	0.32
What: Outcomes[a]	(0.38)	(0.08)
Mean	0.38	0.32

[a] Correlations on this dimension are of questionable meaning because scores showed almost no variation. It has little effect on the mean, however; the disattenuated mean without this dimension is .35.

There is also little variation across dimensions in the correlations between portfolio composites and the Uniform Test, even though one might expect the dimensions to vary in their relationship to the knowledge and skills tapped by a traditional mathematics test.

More ground for pessimism appeared when mathematics portfolio scores were compared to the Uniform Tests in both writing and mathematics. One would expect math portfolio scores to correlate more strongly with the mathematics Uniform Test than with the writing Uniform Test, both for substantive reasons and because of the limited scope of the writing test. Further, one might expect the correlations to vary among dimensions in predictable ways. In the following tables, we present correlations involving two dimensions each in writing and mathematics. In mathematics, we selected Presentation because it was the most reliable of the communications dimensions and Procedures because it seemed most related to the skills that would be needed to solve problems on the mathematics Uniform Test. In writing, we selected Usage, Grammar, and Mechanics because it was the most reliable scoring dimension and because it entails a discrete set of skills (such as punctuation and parallel use of tense) that should be largely unrelated to scores on the mathematics portfolio. We also selected Organization because we reasoned that it might be more similar to some of the skills needed for the math portfolio; poor organization of presentations would lower scores on the mathematics portfolio as well as on the writing test.

Thus, we had a number of expectations in examining the correlations between scores on the math portfolios and the two Uniform Tests. We expected math portfolio scores to be more highly correlated to scores on the math Uniform Test than on the writing Uniform Test. We expected math portfolio scores to have the lowest correlation with Usage. We expected Procedures to have a higher correlation than Presentation with mathematics Uniform Test scores, and we expected Presentation to show higher correlations than Procedures with writing Uniform Test scores.

In the main, these expectations were not borne out. In fourth grade, mathematics portfolio scores (averaged across dimensions) showed nearly identical correlations with writing Organization, writing Usage, and the mathematics Uniform Test (Table 44). In eighth grade, mathematics portfolio

Table 44

Average Disattenuated Correlations Between Math Portfolio Scores and Uniform Test (UT) Scores in Writing and Math

	Writing UT: Organization	Writing UT: Usage	Math UT
Grade 4	.33	.33	.35
Grade 8	.35	.38	.31

scores showed trivially higher correlations with the writing Uniform Test than with the mathematics Uniform Test.

In general, these correlations were similar across the seven mathematics portfolio dimensions, but there were a few dimensions that differed. There was some limited concordance with our expectations in eighth grade. The math portfolio Procedures scores did in fact correlate substantially more strongly with math Uniform Test scores than did the Presentation scores (Table 45). However, any optimism fostered by that pattern is tempered by the fact that the math portfolio Procedures scores correlated nearly as well with both of the writing dimensions as with math Uniform Test scores. (Recall that we expected the correlation between math Procedures and writing Usage to be particularly low.)

Evidence About Program Implementation

Our data on the implementation of the program also cast some doubt on the validity of scores when used for certain purposes. As noted in Chapter 2, teachers report wide variations in their implementation of the program, and some of the variations they report could have a substantial impact on the meaning of scores. For example, differences in rules about revision—how much revision is allowed, how much guidance is provided for revision, and what help is

Table 45

Disattenuated Correlations Between Math Portfolio Scores and UT Scores in Writing and Math, Grade 8, by Dimension

Math dimension	Writing UT: Organization	Writing UT: Usage	Math UT
Presentation	.32	.37	.19
Procedures	.38	.39	.42

allowed from parents and others—could substantially influence scores. Such variations would undermine the validity of comparisons between classes or schools that had substantially different practices. Other variations in implementation that could threaten validity would be differences in the extent of preparation for tasks (and, conversely, their degree of novelty) and differences in the presentation of tasks.

The considerable differences in task assignments we found in our qualitative review of mathematics portfolios could similarly undermine the validity of comparisons among schools. Some tasks afford more opportunity than others to display the scored competencies. Similarly, as some raters have pointed out, one can increase the probability that a student will score well on the portfolio dimensions by assigning tasks that are relatively easy. Even when tasks are nominally the same, teachers can assign easier or more difficult variants.[3]

Conclusions

The evidence presented here, although exploratory and tentative, suggests that the validity of Vermont portfolio scores in mathematics may be questionable. It also illustrates the difficulty of validating assessments of this sort and suggests that researchers will need to cast their nets broadly to get an adequate view of validity.

Our data collection in the 1992-93 years will add additional information relevant to validity. For example, Vermont added a "portfolio-like" task to the mathematics Uniform Test in the spring of 1993, which will provide another useful comparison to portfolio scores. At our request, a subsample of writing portfolios were scored by an alternative method in which scores for each piece in the "rest" were separately recorded, in addition to the "rest" score. We observed criterion sessions (in which raters scored benchmark pieces and then debated their scores) and recorded information on the bases of raters' disagreements. We also obtained feedback from a substantial number of mathematics raters about factors relevant to validity.

---

[3] Raters in the 1993 scoring discussed this problem in some detail and provided concrete examples.

Nonetheless, further expansion of methods and measures will be needed to get a solid understanding of validity. In particular, validation of scores from the Vermont program—and similar programs—will require clarification of the domains that the assessments are designed to measure. This is likely to entail both clearer conceptual definitions and more explicit delineation of the types of tasks and performances that are expected.

CHAPTER 7. IMPLICATIONS

The Vermont program has some unusual features, and some of the findings reported here reflect its idiosyncrasies. However, our findings also have important implications for performance assessment programs more generally and for the design of research to evaluate the quality and effects of those programs.

Vermont's program differs from many current large-scale performance assessment programs in its reliance on portfolios. While most large-scale programs rely primarily on standardized products, such as students' performances on standardized tasks or essays written in response to standardized prompts, Vermont uses standardized performance assessments only as components of its "uniform" tests. The portfolios are unstandardized. The Vermont approach is also atypically "bottom-up." For example, largely volunteer committees of teachers (rather than state Department of Education experts or outside contractors) have much of the responsibility for designing rubrics and establishing guidelines for the form and content of portfolios.

As important as these characteristics may be, they should not obscure the many similarities between the Vermont assessment and other programs or the implications of the Vermont experience for these other efforts. For example, the Vermont program shares with many programs the dual, fundamental goals of measuring student performance and spurring improvements in educational practice. The specific types of instructional change the Vermont program is intended to spark (such as more extensive writing throughout the curriculum and more emphasis on problem solving and communication in mathematics) are also among the primary goals of many other reform efforts across the nation. Moreover, many current proposals call for assessment systems that are similar to the Vermont program. For example, portfolios and other unstandardized products are central to the proposals of the New Standards Project.

For these reasons, the Vermont experience has substantial implications for the performance assessment movement nationwide. The results described here can help set expectations for other programs and provide guidance for their design. In this chapter, we discuss four issues of general importance to

performance assessment that are illuminated by the Vermont experience: expectations regarding the quality of measurement; expectations regarding the impact of assessment on educational practice; the fundamental tension between the goals of educational improvement and measurement quality that motivate this and other performance assessment programs; and requirements for program evaluation.

Expectations for Quality of Measurement

In 1991-92, the Vermont program was largely unsuccessful in providing high-quality information about student achievement. The reliability of scoring was so low that it precluded most of the intended uses of the portfolio scores. Moreover, both patterns in the scores themselves and variations in the program's implementation raise doubts about whether the scores would have provided a valid basis for certain conclusions—among them, comparisons across schools or other groups—even if the scoring had been more reliable.

A key question for policy is *why* the assessment data were so weak. For example, members of the Vermont State Board of Education wanted to know how much of the unreliability of scoring is a consequence of using portfolios and how much improvement in reliability could realistically be expected if the program was improved but continued to rely on portfolios. Similarly, observers outside of Vermont want to know how much the problems documented here can be attributed to factors that will affect their own programs.

Our view is that the problems encountered in Vermont should serve as a signal to set modest expectations for the quality of data from innovative performance assessments, particularly over the short- and moderate-term. Although the problems in Vermont stem partly from idiosyncratic factors, they also appear to reflect factors relevant to many performance assessment programs.

Our observations of the program suggest at least three possible causes of the unreliability of scoring—problems with the scoring rubrics, insufficient training, and the lack of standardization of tasks—but we lack the data at this time to disentangle their relative contributions. Although all three have aspects that are unique to Vermont, it is likely nonetheless that similar problems will arise in other programs, particularly those that rely on portfolios or other types

of unstandardized performance assessments. For example, reliance on nonstandardized tasks, however desirable for possible effects on instruction, will often severely complicate efforts to devise reliable scoring rubrics and methods, particularly in subjects (such as mathematics and science) in which tasks are likely to vary greatly.[1] Similarly, the task of training large numbers of teachers to score reliably will generally be difficult in most programs and will be especially hard in programs that require teachers to grade disparate products using methods not tailored to any specific tasks.

Similarly, the problems of validity suggested by our data most likely stem at least in part from factors common to many assessment programs. For example, our generalizability analysis of mathematics portfolio scores showed large task-to-task variations in the scores of individual students that would threaten the validity of inferences about student performance based on a small number of tasks. Far from being unique, this limited generalizability of performance across complex tasks within a subject area is the norm in the research on performance assessments (see, for example, Dunbar, Koretz, and Hoover, 1991, and Shavelson, Baxter, and Gao, 1993). The large variability we found in teachers' implementation of the Vermont program is also likely to be mirrored in other programs that attempt to integrate assessment into teacher-directed instruction, and it will pose potential threats to validity in those programs as well.

Whatever its causes, the *effects* of unreliable scoring discussed in the preceding chapters are not unique to Vermont. The appropriate uses of scores will be limited whenever similar problems of reliability arise. Unreliable scoring will of course always undermine the utility of scores for making decisions about individual students. Moreover, as the results above illustrate, it will also threaten inferences about aggregates. For example, unreliability of scoring will generally bias the distribution of scores, causing too many students to score at the extremes and too few near the middle. Thus, estimates of the proportion of

---

[1] We are aware of one portfolio program in which interrater agreement was far higher than in Vermont. A recent portfolio assessment of writing in Pittsburgh achieved interrater correlations above .70 (LeMahieu, 1992). We suspect that one reason Pittsburgh attained reliability so much higher than that of the Vermont writing portfolio program is that in the Pittsburgh program, portfolios were scored by a relatively small group of people who had long involvement in the program. We are not aware, however, of any large-scale program that has achieved comparably high levels of agreement in mathematics or science portfolio assessments in which the contents of portfolios are as unregulated as they have been in Vermont.

students reaching various points on the scale will be misleading overall, and comparisons between groups (schools, districts, demographic groups, or whatever) that differ substantially in their average scores will be error-prone. Similarly, although complex analytical scoring systems may be beneficial as incentives for instructional change, they will not yield meaningful data if the various scoring dimensions cannot be distinguished reliably by raters. It is true that these problems were extreme in Vermont in the 1991-92 program, but they would remain substantial even with considerable improvements in reliability. For example, even if reliability at the level of scoring dimensions was increased to .70—which would represent a very large improvement—fully half of the variance in students' scores would still be error, and that much error would substantially bias the proportion of students reaching each score.

Expectations for Impact on Educational Practice

At least in mathematics, the Vermont assessment program appeared to be more successful in 1991-92 as an educational intervention than as a measurement program. Principals and teachers agreed that the program provided a powerful impetus for change in instruction, and the reported changes, such as an increased emphasis on problem solving, appeared to be largely consonant with the goals of the program. The fact that so many schools opted to expand their use of portfolios beyond the fourth and eighth grades despite the large burden it imposed is a telling measure of its perceived positive effects on instruction.

Many observers—we among them—see these preliminary findings as grounds for optimism about the potential effects of innovative assessments on instructional quality. There are, however, reasons to temper that optimism. The evidence to date about the effects of the program is both limited and mixed, and the Vermont experience underscores how difficult it is to obtain desired outcomes.

It is important to reiterate some of the most important limitations of the data reported here. The information we report on the effects of the program reflect primarily self-reports: interviews with principals and teachers and an anonymous teacher questionnaire. Our qualitative analysis of portfolios was limited in scope and not necessarily representative of the state as a whole. Our direct observation of classrooms was extensive in terms of sampling but very

limited in duration and depth; although it provided useful examples and clarified and supplemented some of the responses we obtained from educators, it was not sufficient to provide a systematic check on the accuracy of self-reports. Moreover, the ultimate test of instructional improvement is enhanced learning. The data we had from the 1991-92 implementation offer no direct measure of effects on student learning.

If one accepts the reports of teachers and principals as an indication of positive effects on instruction, there are still reasons to be cautious about the extent and pervasiveness of that impact. One reason is the patterns shown by the portfolio scores themselves. The unreliability of scoring suggests inconsistent interpretation of performance goals by the state's teachers, and that in turn raises the prospect of inconsistent instructional goals and practices. The apparent lack of independence of most of the scoring dimensions and the relatively minor differences between the best-piece and rest scores in writing indicate that on average, teachers would learn no more from the many scores assigned to each portfolio than they would from a single score. This raises the question of whether teachers are in fact providing students with reasonable and consistent feedback on their efforts to meet the performance goals reflected in the assessment's many dimensions and components. (It is possible, however, that requiring scores on multiple dimensions caused teachers to focus instruction on all of them, even if the scores were not a reasonable base for monitoring their efforts.)

In addition, while the Vermont assessment program was apparently a powerful method of signaling to teachers what was expected of them, the evidence suggests that its success in this respect was incomplete. Indeed, the Vermont experience argues that much more than an assessment is needed to accomplish this goal. Our questionnaires and other observations show that in mathematics, the Vermont program, including the provision of illustrative tasks in the *Resource Book* and considerable training, had apparently substantially altered teaching in the aggregate. However, it had not been sufficient to create a consistent understanding of what constitutes appropriate teaching. This is not an entirely negative finding; some observers see the spirited debate among Vermont teachers about curriculum and instruction that continues even two years after the inception of the program to be one of its greatest benefits. Nonetheless, this is one more instance in which the Vermont experience suggests

moderate expectations. It is one thing to communicate to teachers that they should put more emphasis on problem solving; it is quite another to communicate effectively what that means and how it can be accomplished in a way that actually improves students' skills. To do so is likely to require a great deal of time and effort.

This points to yet another reason for caution: the major costs of the progress made to date. There has been no accounting of the direct and indirect financial costs of the system; indeed, given the extraordinarily decentralized nature of the Vermont educational system and of responsibility for this program, it would be difficult to obtain one. It is clear, however, that the costs in time, effort, and stress have been large. Indeed, the burdens noted in this report represent only the initial stages of a continuing and still difficult and costly process of program development.

Finally, one has to ask here the same question we asked about quality of measurement: To what extent do our findings reflect the idiosyncrasies of the Vermont program? It is our impression that the answer is again mixed. The nature of the assessment tasks themselves, particularly in mathematics, clearly did to some degree signal concretely to teachers what was meant by otherwise abstract goals such as "increasing the emphasis on problem solving." One would expect that this signaling function could be served by diverse performance assessment programs quite unlike the Vermont program. However, it is our impression that the impact of the program also has stemmed in part from aspects of the program that are relatively unusual, albeit replicable. For example, the support of teachers and principals appears critical to the effective operation of the program, and our interactions with Vermont educators suggests that both the decentralized, bottom-up nature of the program and extensive and time-consuming efforts by the state may have been critical in building that support.

Tensions Between the Goals of Assessment Programs

To what extent can assessment programs be expected to meet the dual goals of improving instruction and providing high-quality information about student achievement? Given that both goals are fundamental to the current performance assessment movement, the answer will have widespread ramifications for education reform. It is not surprising that the Vermont program appeared to be

considerably more effective in meeting one goal than the other during its first year of its implementation. More important are implications of the Vermont experience for the longer-term potential of meeting both goals.

Our view is that the goal of improving instruction often conflicts with the goal of providing high-quality, valid, and reliable data about student performance. More concretely, an assessment program designed primarily to meet the first of these goals would likely be quite different from one designed primarily to meet the second. For example, standardization of tasks and administrative conditions will generally improve the quality (at least, the comparability) of data about student performance, but those same attributes are likely to impede the integration of assessment and curriculum and may undermine teachers' feelings of ownership and commitment to the program. For programs (such as Vermont's) that have both goals, success will depend on finding a workable compromise between the two, deciding, for example, what price in measurement quality is acceptable to gain additional leverage on instruction. The founders of the Vermont program, unlike many other reformers, openly confronted this dilemma at the outset, but experience is beginning to show how difficult it will be to resolve it.

The tension between the goals of the program became apparent in a variety of contexts in Vermont. For example, one consideration that led us (and some teachers) to question the validity of comparisons based on portfolio scores was the large variation in key aspects of program implementation, such as policies toward the revision of students' work. Clearly, such variations can threaten validity. Two students of similar competence might produce comparable products when confronted with the same constraints but dramatically different products if one is allowed much more time to revise or is given more help (by teachers, parents, or other students) in revising. Yet some of the variations in instruction that could undermine the validity of comparisons may be precisely those one wishes to encourage to improve instruction. An effective teacher may decide, for example, that less able students need more help than more able students, and perhaps more structured, directive help, in revising products. The teacher may believe, for example, that the more able students are at a point where they should learn to work more autonomously in revising their work. This variation in procedures may help both groups of students, but it will undermine the validity of comparisons based on the scores by making differences

between the groups appear smaller than they really are. Conversely, other differences in revision rules might exaggerate differences in competence.

The tension between the goals of measurement and instructional improvement also arose in developing procedures for scoring portfolios. The goal in Vermont has been to involve all teachers in the affected grades in scoring portfolios. This policy stems directly from the instructional improvement goals of the program: Training in scoring student work is seen as a critical component of training teachers to understand the instructional goals of the program. Yet, the more broadly responsibility for scoring is shared, the more difficult it becomes to provide enough training to bring scorers to an acceptable level of proficiency, and the more likely it becomes that insufficiently proficient raters will participate in the scoring process.[2]

Requirements for Evaluation

Many performance assessment programs, including Vermont's, are intended to have pervasive effects throughout the educational system. These programs, however, like more traditional forms of test-based accountability, are not self-evaluating. Upward trends in scores on the new measures are to be expected and are not sufficient to indicate that the goals of the programs are being met. The range of questions that need to be investigated to evaluate such a program is illustrated by our experience in Vermont—both by the findings noted above and by the many questions our data do not address.

Documenting Program Implementation

The findings described in Chapter 2 illustrate the importance of documenting not only typical patterns of implementation but also variations among teachers, schools, and categories of students. This information is important for formative purposes—that is, to identify problems that need to be addressed as the program matures. Variations in implementation also may have

---

[2] In response to this conflict in goals, we suggested to the Vermont State Department of Education that scoring be conducted on two separate tracks. The Department would continue to provide training in the scoring process to all teachers and would continue to request that all teachers score. However, scores for reporting by the Department would be generated separately at a single workshop at which raters would receive additional training and would be monitored. The Department followed this suggestion in the 1992-93 school year, and the effects of this change and further maturation of the program on the quality of scores will be discussed in a forthcoming report.

important implications for equity. A preliminary investigation of variations in principals' responses to our interviews did not reveal striking differences between large and small schools or between high-poverty and other schools. Nonetheless, it seems likely that in other contexts, variations in program implementation (and quality) may be associated with factors such as socioeconomic status or ethnicity.[3] As we noted earlier, variations in how programs are implemented may also have substantial implications for the validity of the assessment results. Information on differences in implementation can alert users to this possibility and can be helpful for designing validation studies.

The Vermont experience also underscores the importance of a far-ranging investigation of costs and burdens. Particularly when assessments require substantial efforts by classroom teachers, it is simply inadequate to tabulate only the direct costs borne by states or large districts. The nonfinancial costs are diffused throughout the educational system, and financial costs may be hidden in other budget categories at lower levels of the system. (One example of this that we found in Vermont was the substantial allocation from districts' substitute budgets to pay for release time for training.) Proponents of systemic reforms based on performance assessment often maintain that some of the burdens imposed by the assessment program should not properly be considered purely costs of assessment. For example, they argue that some of the time teachers spend adapting to the systems should be treated partly as costs of professional development or curriculum improvement. The Vermont experience is consistent with this view; for example, the large amount of time some teachers devoted to finding appropriate tasks can clearly be seen as curriculum improvement as well as a cost of assessment. However, this is an argument for more complete investigation of costs and burdens, not less.

Investigating Instructional Effects

Our experience suggests that more direct measures of instructional change would be very valuable. However, they may be difficult to obtain. Evaluators

---

[3] Vermont has a substantial poor population, but it lacks many of the other social divisions that are of concern to other jurisdictions. In addition, the mechanisms by which teachers sort themselves among schools may be very different in Vermont (where most districts operate only a few schools and are geographically dispersed) than in jurisdictions with large districts and geographically concentrated schools.

should ideally obtain baseline measures before the program is implemented, and the current political climate—exacerbated as it has been by widespread, unrealistic expectations about the speed with which reforms of this sort can be effected—may make it very difficult to put data collection into place before programs are fielded. Moreover, in large-scale and geographically dispersed assessment programs, the costs of some direct measures may be prohibitive. Other research on curriculum—for example, the continuing efforts to develop more sensitive measures of curriculum for international studies of achievement—may provide measures that are less burdensome than direct observation but that are still useful for evaluating programs of this kind.

Assessing Reliability

Estimates of interrater agreement are necessary but clearly insufficient. Evaluators need to explore both the causes and the effects of differences among raters. Their causes will often prove more difficult to ascertain in programs that, like Vermont's, rely on unstandardized products and general-purpose scoring rubrics. For example, our data yielded estimates of the variability of performance across tasks in mathematics but did not provide any information about the impact of specific task characteristics because tasks are not specifically identified.[4] In standardized performance assessments, by contrast, aspects of tasks can be systematically varied. Unstandardized assessments also introduce large but largely undocumented variations in task administration. Evaluators may be able to obtain some of the needed information about this by means of interviews and questionnaires, but it may also be necessary to introduce planned variations into the operation of the assessment program to evaluate the quality of scores adequately.[5]

Measuring Student Performance

It is perhaps ironic that one of the most difficult problems in evaluating programs of this sort, which are themselves designed to measure student performance, is obtaining adequate measures of student performance.

---

[4] In addition, teachers will often use different variants of common problems. For example, mathematics teachers in the 1993 scoring pointed out variants of a single problem used by different teachers that differed markedly in difficulty.

[5] In Vermont, the open-ended questions in the uniform assessment should ideally provide some supplementary information. We have also proposed introducing standardized tasks (albeit with less than fully standardized administration) into the operation of the portfolio system itself.

One reason that direct measures of student learning are needed is to gauge positive effects on learning. The new programs are typically intended at least in part to measure things than are not well tapped by extant assessments. Thus, even in jurisdictions that, unlike Vermont, have ongoing assessments that will continue after the inception of the new program, there is a real possibility that trends on the old measure will fail to register positive effects of the new program. (Moreover, scores on the old tests sometimes are not trustworthy as measures of trends on the skills they are supposed to assess. If the new programs lessen inappropriate teaching to the test on the old assessments, some decline in scores on the old test may be nothing more than the elimination of bias.)

Additional direct measures of student performance are also critically important for validation. They are needed to test the generalizability of performance even when the new programs are in their infancy, and they will become only more important over time as the possibility of inappropriate teaching to the test (and inappropriate administration) raises the specter of inflated scores on the new assessments.

Unfortunately, there are many obstacles to obtaining sufficient direct measures of student performance. Here again, the pace of reform makes it difficult to obtain baseline measures. The costs of developing and administering the measures will also be an obstacle, as will access to schools (which are often understandably reluctant to allocate yet more time to testing). In addition, in Vermont, one of the most serious hindrances to independent measurement of student performance has been the insufficient delineation of the domains that the new assessments are supposed to measure. As performance assessments in subjects other than writing become more common, we expect this problem to arise in many other programs as well.

Conclusions

The experience of the Vermont portfolio program to date suggests the need for moderate expectations, patience, and ongoing evaluation, not only in Vermont, but in other performance assessment programs as well. As Richard Mills and Ross Brewer acknowledged at the outset, the Vermont program (and, we add, many other performance assessment programs) will require a long period of development (Mills & Brewer, 1988). Perhaps even more important,

the Vermont experience illustrates the tensions among the goals of this and similar programs and the need to make difficult trade-offs in mediating among them. Only time and careful scrutiny will show how fully the goals of the Vermont program—and of similar reform programs centered on performance assessment—can be met, as well as what steps will need to be taken to meet them.

BIBLIOGRAPHY

Cochran, W.G. (1977). ***Sampling techniques*** **(3rd ed.). New York: Wiley.**

Cronbach, L.J., Gleser, G.G., Nanda, H., & Rajaratnam, N. (1972) ***The dependability of behavioral measurements: Theory of generalizability for scores and profiles.*** **New York: John Wiley and Sons.**

Dunbar, S., Koretz, D., & Hoover, H.D. (1991). Quality control in the development and use of performance assessment. ***Applied Measurement in Education, 4*****(4), 289-303.**

Hartley, H.O., Rao, J.N.K., & LaMotte, L.R. (1978). A simple synthesis based method of estimating variance components. ***Biometrics, 34,*** **233-243.**

Herman, J.L., Gearhart, M., & Baker, E.L. (in press, 1993). Assessing writing portfolios: Issues in the validity and meaning of scores. ***Educational Assessment.***

Hieronymus, A.N., Hoover, H.D., Cantor, N.K, & Oberly, K.R. (1987). ***Writing: Teacher's guide.*** **Chicago: Riverside.**

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992, December). ***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., & Deibert, E. (1992). ***The Vermont portfolio assessment program: Interim report on implementation and impact, 1991-92 school year*** **(CSE Tech. Rep. No. 350). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.**

LeMahieu, P. (1992, April). Data from the Pittsburgh writing portfolio assessment. In J. Herman (Chair), ***Portfolio assessment meets the reality of data.*** **Symposium presented at the annual meeting of the American Educational Research Association, Atlanta.**

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.**

Salmon-Cox, L. (1982, September). ***MAP Math: End of year one report.*** **Unpublished manuscript, University of Pittsburgh, Learning Research and Development Center.**

Salmon-Cox, L. (1984, September). ***MAP Reading: End-of-year report.*** **Unpublished manuscript, University of Pittsburgh, Learning Research and Development Center.**

Shavelson, R.J., & Webb, N.W. (1991). ***Generalizability analysis.*** **Newbury Park: Sage.**

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. ***Journal of Educational Measurement, 30*****(3), 215-232.**

Vermont State Department of Education. (1990, September). ***Vermont writing assessment: The pilot year.*** **Montpelier: Author.**

Vermont State Department of Education (1991a). ***Looking beyond "The Answer:" The report of Vermont's mathematics portfolio assessment program.*** **Montpelier: Author. [undated]**

Vermont State Department of Education. (1991b). ***"This is my Best:" Vermont's writing assessment program, pilot year 1990-91.*** **Montpelier: Author. [undated]**

Vermont State Department of Education. (1991c, September). ***Vermont mathematics portfolio project: Teacher's guide.*** **Montpelier: Author**

Vermont State Department of Education. (1991d, September). ***Vermont mathematics portfolio resource book.*** **Montpelier: Author.**

APPENDIX A. RATER BIASES

In both fourth and eighth grades, each writing portfolio was first rated by the student's own teacher; a sample was then re-scored by an independent rater. To explore any biases introduced by using the student's classroom teacher to rate his or her portfolio, we compared the mean score given by the classroom teacher to those given by the independent rater for the sample of portfolios scored twice. This comparison was made for both components of the portfolio (the best piece and the "rest") on each of the five scoring dimensions.

The differences in means were generally very small, in most cases too small to be of any practical importance. For example, we found that in Grade 4, the mean score assigned to best pieces by students' own teachers on the "Purpose" dimension was 2.98; the corresponding mean for independent raters was 2.93. With one exception, all of the other differences were less than 0.1. The means for the remaining components are given in Table A.1 for Grade 4 and Table A.2 for Grade 8.

With the exception of a single comparison (Voice/Tone for fourth-grade best pieces), none of the differences were statistically significant (using a critical level of .05 and not adjusting for multiple comparisons). T-statistics for all differences are shown in Tables A.1 and A.2. The estimated mean scores and standard deviations were found using a jackknife technique (Cochran, 1977). Jackknife estimates were used because students were sampled from schools. This clustering among students within schools may create a correlation among student scores from the same school. (That is, students from within a school may be more similar than randomly selected students.) The jackknife estimates make the appropriate adjustment to standard errors to account for this clustering.

Also, raters could mark a portfolio or a piece of the portfolio as "non-scorable," creating missing data for that case. If one rater identified a portfolio or piece as non-scorable but the other rater did not, the available score was used in estimating the mean. Thus, a given portfolio could enter into one mean (the classroom teacher or independent rater) but not the other. Our results were not sensitive to this decision. Additional estimates were calculated by eliminating from both means all pieces identified as non-scorable by either rater, and these were the same as the means we used to two decimal places.

Table A.1

Mean Scores for the Two Ratings, Grade 4

		Student's teacher		Independent rater			
Piece	Dimension	Mean	Std Error	Mean	Std Error	Mean Diff	T[a]
Best	Purpose	2.98	0.025	2.93	0.026	0.05	1.58
Best	Organization	2.90	0.026	2.88	0.027	0.01	0.47
Best	Details	2.81	0.028	2.76	0.025	0.05	1.47
Best	Voice/Tone	2.77	0.033	2.67	0.026	0.10	2.73
Best	Usage/Grammar /Mechanics	2.84	0.027	2.87	0.026	-0.03	-1.15
Rest	Purpose	2.85	0.026	2.79	0.029	0.06	1.78
Rest	Organization	2.74	0.027	2.71	0.030	0.04	1.02
Rest	Details	2.59	0.029	2.57	0.031	0.02	0.57
Rest	Voice/Tone	2.60	0.032	2.54	0.028	0.07	1.79
Rest	Usage/Grammar /Mechanics	2.71	0.028	2.70	0.028	0.01	0.48

[a] T gives the ratio of the difference in means to the estimated standard error of the difference. This statistic is approximately normal and can be compared to a Standard Normal Table.

Table A.2

Mean Scores for the Two Ratings, Grade 8

		Student's teacher		Independent rater			
Piece	Dimension	Mean	Std Error	Mean	Std Error	Mean Diff	T[a]
Best	Purpose	3.13	0.038	3.08	0.040	0.05	1.18
Best	Organization	3.04	0.038	2.98	0.037	0.06	1.56
Best	Details	2.95	0.043	2.95	0.042	0.00	0.04
Best	Voice/Tone	3.00	0.045	2.96	0.042	0.03	0.85
Best	Usage/Grammar /Mechanics	2.83	0.039	2.82	0.039	0.01	0.27
Rest	Purpose	2.92	0.037	2.93	0.035	-0.01	-0.27
Rest	Organization	2.82	0.043	2.80	0.040	0.02	0.48
Rest	Details	2.70	0.039	2.73	0.037	-0.03	-0.73
Rest	Voice/Tone	2.79	0.043	2.74	0.043	0.04	1.04
Rest	Usage/Grammar /Mechanics	2.63	0.036	2.65	0.036	-0.02	-0.57

[a] T gives the ratio of the difference in means to the estimated standard error of the difference. This statistic is approximately normal and can be compared to a Standard Normal Table.

APPENDIX B

Generalizability analysis investigates the dependability of person's score on a single measure of ability. A dependable measure must generalize from a person's observed score to the average score the person would receive on all acceptable tests (Shavelson & Webb, 1991). A given measure, for example, a student's score from a single rater on single part of a portfolio, represents performance on single realization of many alterable conditions. For example, the student could have produced a different part for the portfolio or a different rater could have scored the portfolio. Few people are interested in a student's performance in a specific instance; rather, they are interested in the extent to which the student possesses the attribute that instance of measurement purports to assess. Thus, the meaningful score is not the student's score from a given rater on a given part but the average over all scores the student would receive from all similar raters and similar parts. The score for given specific part and rater must generalize to this average over all parts and raters. These factors that one plans to generalize over are called facets (Shavelson & Webb, 1991).

In the Vermont writing portfolios, each score depends on three factors, the student and the two generalizable facets, rater and part. Thus the score can be modeled as

$$
\begin{aligned}
y_{ijk} = {} & \mu \\
& + \mu_i - \mu && \text{(person effect)} \\
& + \mu_j - \mu && \text{(rater effect)} \\
& + \mu_k - \mu && \text{(part effect)} \\
& + \mu_{ij} - \mu_i - \mu_j + \mu && \text{(person by rater effect)} \\
& + \mu_{ik} - \mu_i - \mu_k + \mu && \text{(person by part effect)} \\
& + \mu_{jk} - \mu_j - \mu_k + \mu && \text{(rater by part effect)} \\
& + y_{ijk} - \mu_{ij} - \mu_{ik} - \mu_{jk} \\
& \quad + \mu_i + \mu_j + \mu_k - \mu && \text{(residual effect)}
\end{aligned}
$$

where i = 1 to the number of students (1741 for Grade 4 and 655 for Grade 8), j = 1 to the number of raters and k = 1 (best piece) or 2 (rest). Only the subset of student portfolios whose raters could be identified were included in this analysis.

The procedure used for scoring student portfolios did not produce a fully crossed data set. Each portfolio was scored by only two raters, not by all raters (400 in Grade 4 and 162 in Grade 8). This creates an unbalanced data set.

In this study parts are "fixed." That is, we are interested in generalizing to any portfolio constructed with a best piece and the rest. Because part is fixed, no variance component will be estimated for it. The effects of students' specific selection of pieces to be included as best or rest contribute to the student-by-part interaction. That is, the student-by-part interaction measures variability among the part scores from a single student. This variability may be the result of the student simply performing better on one of the parts or from the student's specific selection of pieces for both parts. These two sources of variability cannot be disentangled in our analyses.

Our analysis only measures the extent of student item selection to the extent that the full range of possible pieces is represented in the items chosen to be in the portfolio. However, this is probably not a representative sample. Students were instructed to select their best work to be included in the portfolio. Thus, if Josh writes stories poorly, he may not have selected any stories for his portfolio. His portfolio scores would therefore not reflect the variability that would have been found if his portfolio had included stories. On the other hand, if Josh is weak in writing stories he would most likely never include a story in any portfolio he might construct. Therefore, the variability found in these portfolio scores may be representative of the variability that one would expect to find in scores from student-selected portfolio pieces.

Generalizability theory estimates the generalizability of a score by using the observed scores to estimate the variability of each effect or facet. For example, for the writing portfolios, we estimated the variability among different raters and the various interaction effects. We also estimated the variability among students. Student variability is considered the true measure of the variation in student ability. This variance is an estimate of the variability that exists in students' "universe" scores—the mean score a student would receive over all possible two-part self-selected portfolios scored using the current scoring procedure by all possible similar raters. The interpretation of the universe score is not clarified in generalizability analyses; it must be inferred from other sources such as validity studies.

For each effect given in the above model, excluding the fixed part effect, we estimated the variability of that effect in the population of all such effects. These estimates indicate the amount of the variance in all possible portfolio part scores by all similar raters that is attributable to each effect. All effects other than the student effect are noise. That is, these effects are the results of the specific portfolio and rater and are not associated with the student's universe score.

Because of the unbalanced nature of the sample of scores, the traditional ANOVA-based estimates of components of variance were not available. Furthermore, the large sample size made it infeasible to use other ANOVA-based methods to estimate the variance components. The component estimates were found using the MIVQUE estimation procedure (Hartley, Rao, & LaMotte, 1978). This method produces unbiased, (locally) minimum variance estimates of the variance components.

The estimates of the variances for the effects given in our model for student scores are in Table B.1 for Grade 4 and Table B.2 for Grade 8. The estimates were found separately for each dimension.

The generalizability of the portfolio score for a single dimension is measured using the generalizability coefficient (Shavelson & Webb, 1991). The generalizability coefficient is approximately equal to the expected value (average) of the square of the correlation between the observed scores and the student's universe scores (Shavelson & Webb, 1991). It is also approximately equal to the correlation between observed scores on two analogous portfolios. For example, the generalizability coefficients given in Tables B.1 and B.2 are estimates of the correlation between two portfolios selected by the same student, each scored by a single rater who reads all pieces and assigns one score to the best piece and one score to the rest. This differs from the correlation coefficients given in Chapter 3 because this coefficient measures the degree of agreement between similar portfolios and raters simultaneously, rather than measuring agreement between raters on the same portfolio.

The advantage of this generalizability study is that because we have estimated the components of variance, we can estimate the expected square of the correlation between universe scores and observed scores for various portfolio plans and scoring schemes. For example, we can estimate this correlation for a

Table B.1

Grade 4 Writing Variance Component Estimates

	Purpose		Organization		Details		Voice/Tone		Usage/ Grammar/ Mechanics	
Source	Est.	%	Est.	%	Est.	%	Est.	%	Est.	%
Student	0.147	27.5	0.163	28.7	0.152	26.5	0.158	23.8	0.195	32.7
Student by Parts	0.032	6.0	0.040	7.0	0.042	7.4	0.042	6.3	0.056	9.4
Rater	0.069	12.9	0.065	11.5	0.077	13.5	0.111	16.7	0.052	8.8
Rater by Student	0.007	1.3	0.006	1.1	0.009	1.7	0.007	1.1	0.006	1.1
Rater by Parts	0.085	16.0	0.090	15.9	0.110	19.1	0.137	20.6	0.115	19.2
Residual	0.193	36.2	0.204	35.9	0.183	31.9	0.209	31.4	0.172	28.8
Gen. Coef.	0.43		0.43		0.41		0.37		0.46	

Table B.2

Grade 8 Writing Variance Component Estimates

	Purpose		Organization		Details		Voice/Tone		Usage/ Grammar/ Mechanics	
Source	Est.	%	Est.	%	Est.	%	Est.	%	Est.	%
Student	0.166	31.1	0.195	33.1	0.214	35.4	0.218	33.3	0.281	44.4
Student by Parts	0.022	4.2	0.043	7.2	0.043	7.2	0.047	7.2	0.041	6.4
Rater	0.047	8.9	0.047	7.9	0.046	7.6	0.066	10.1	0.039	6.1
Rater by Student	0.008	1.4	0.008	1.4	0.004	0.7	0.003	0.4	0.006	1.0
Rater by Parts	0.107	20.0	0.122	20.6	0.103	17.1	0.099	15.1	0.104	16.4
Residual	0.184	34.4	0.176	29.7	0.194	32.1	0.221	33.9	0.163	25.7
Gen. Coef.	0.44		0.46		0.49		0.48		0.58	

portfolio with three fixed parts, scored by two raters, where the score is the average of these two scores. The score from a portfolio is assumed to be the total (or average score) over all parts and raters for each dimension.

These generalizability coefficients are meaningful only if we assume that the scoring procedure remains constant. That is, each rater scores all pieces together and then assigns part scores; the independent raters do not each score the parts separately or independently. We must assume the same scoring procedure because changing the scoring procedure might change the variability among scores. Also, this generalizability coefficient is only meaningful if we are considering similar parts constructed of similarly selected pieces.

Under these assumptions, we can estimate the generalizability coefficient for any combination of parts and raters. The generalizability coefficients are calculated using the following formula

$$\frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_{sr}^2 / n_r + \hat{\sigma}_{sp}^2 / n_p + \hat{\sigma}_{srp.e}^2 / n_r n_p}$$

where $\hat{\sigma}_i^2$ i = s, sr, sp and srp.e denote the estimated variance components for students, student by rater, student by part and residual effects respectively, and n_r and n_p denote the number raters scoring each portfolio and the number of parts respectively.

Typically the adjusted student-by-part effect would be included in the numerator as well as the denominator for fixed parts. However, although parts are fixed (best or rest), the pieces the students include in each part would change if the student were to construct an alternative portfolio. Thus the student-by-part interaction varies with portfolio and scoring, and this term belongs only in the denominator. For Grade 4, Figures B.1 to B.5 show the effects on the coefficient from varying the number of parts and the number of raters. Figures B.6 to B.10 are the analogous plots for Grade 8.

Figure B.1. Effects on Reliability of Increasing the Number of Raters or Pieces,

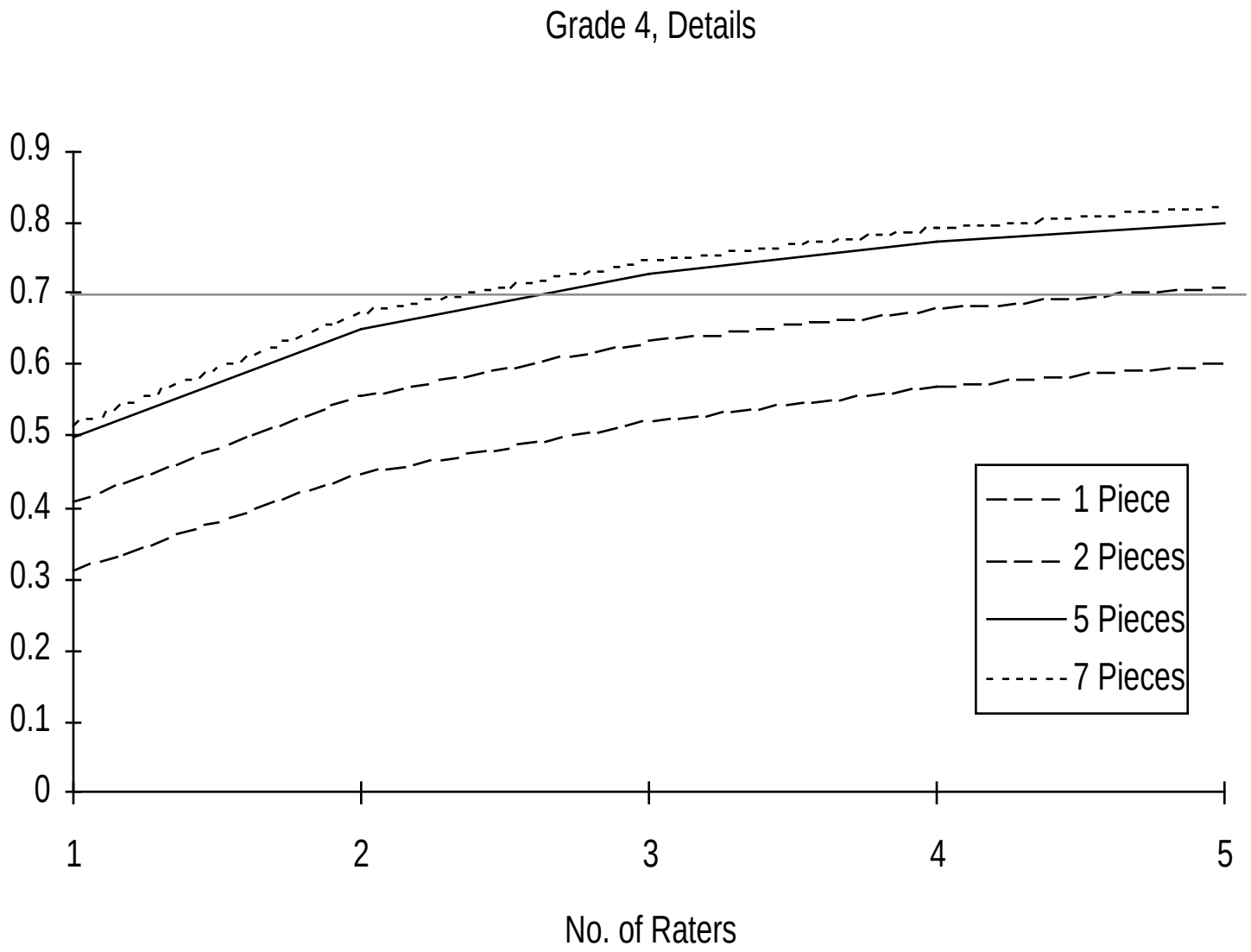


Figure B.2. Effects on Reliability of Increasing the Number of Raters or Pieces,

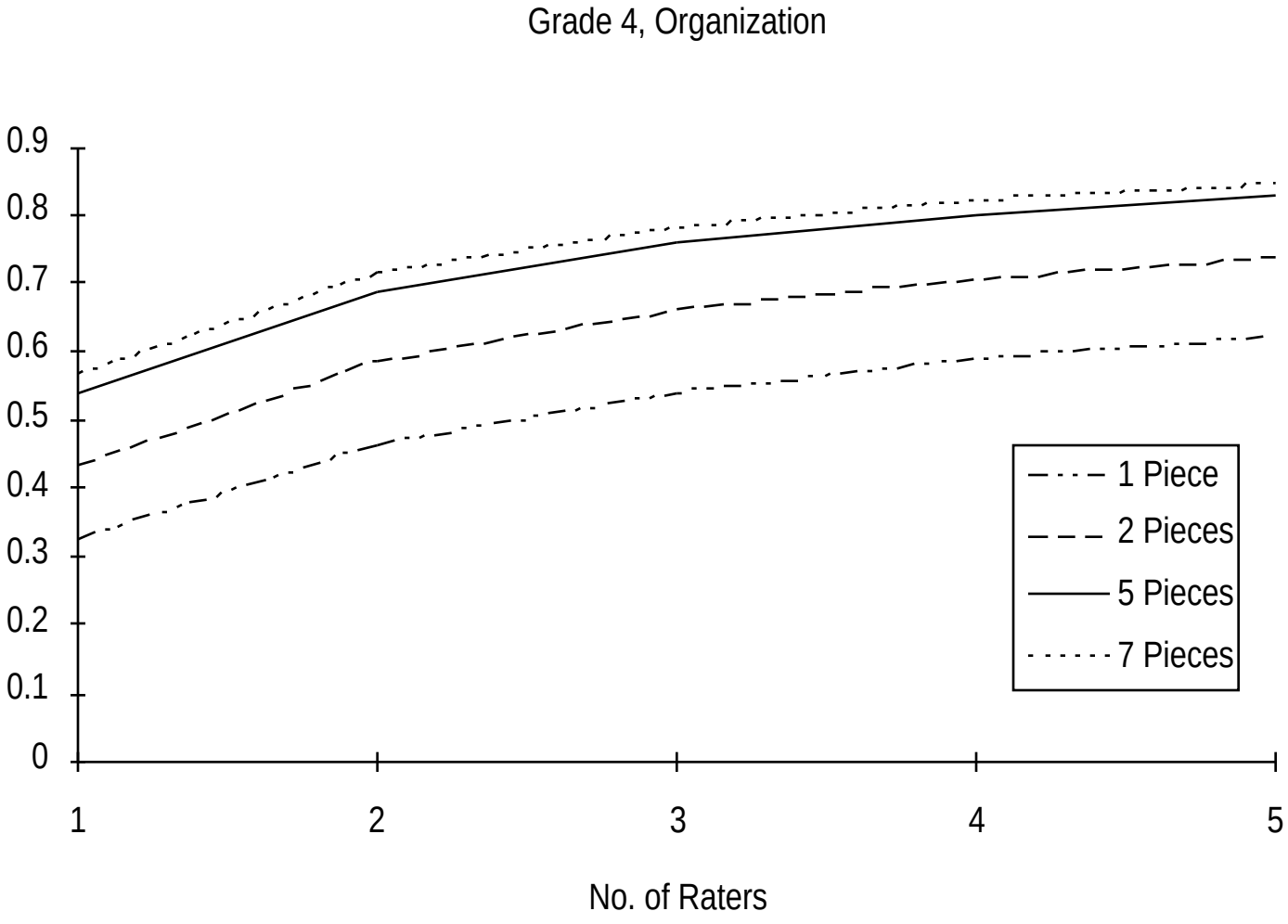


Figure B.3. Effects on Reliability of Increasing the Number of Raters or Pieces,

Grade 4, Purpose

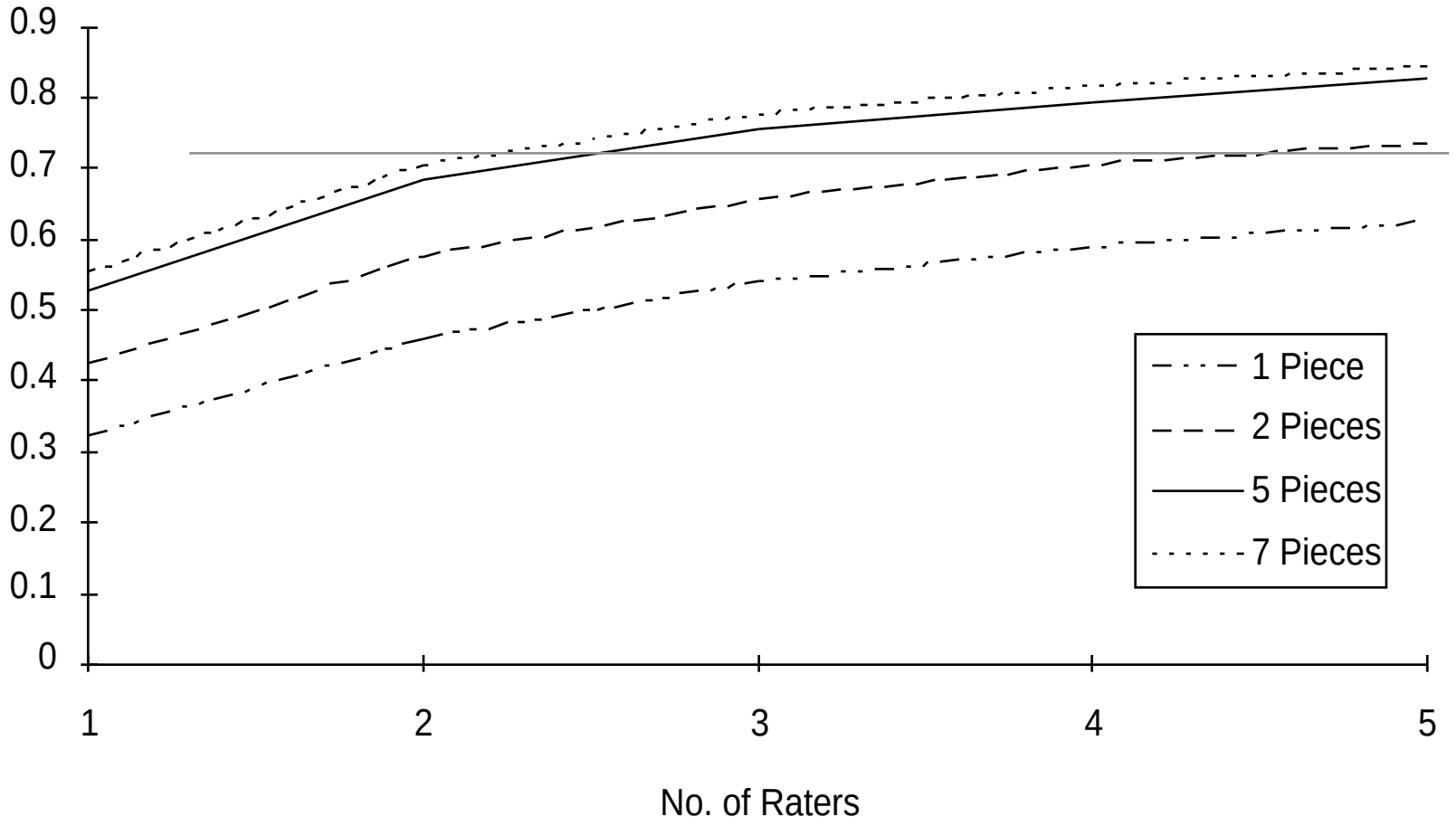


Figure B.4. Effects on Reliability of Increasing the Number of Raters or Pieces,

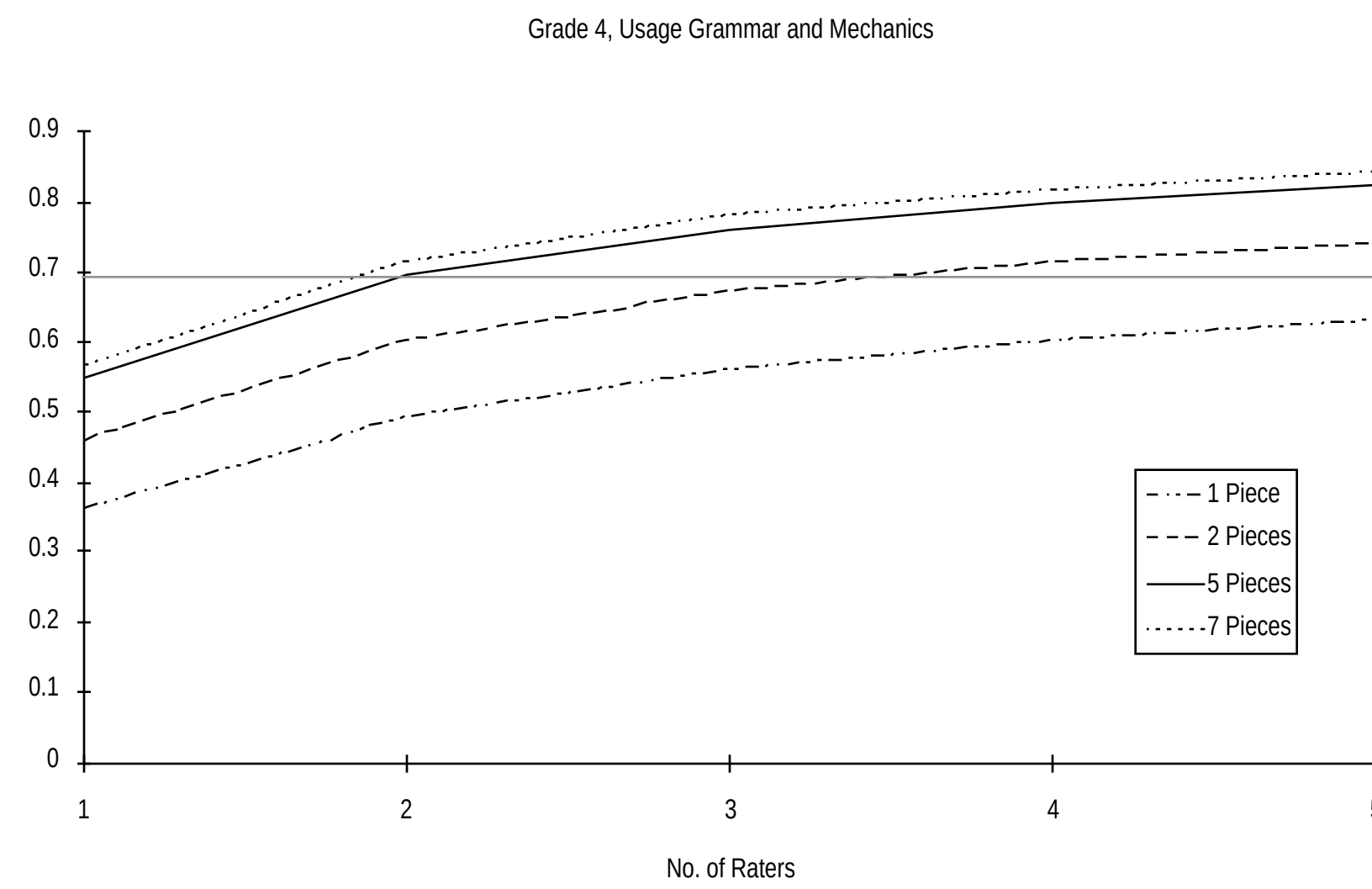


Figure B.5. Effects on Reliability of Increasing the Number of Raters or Pieces,

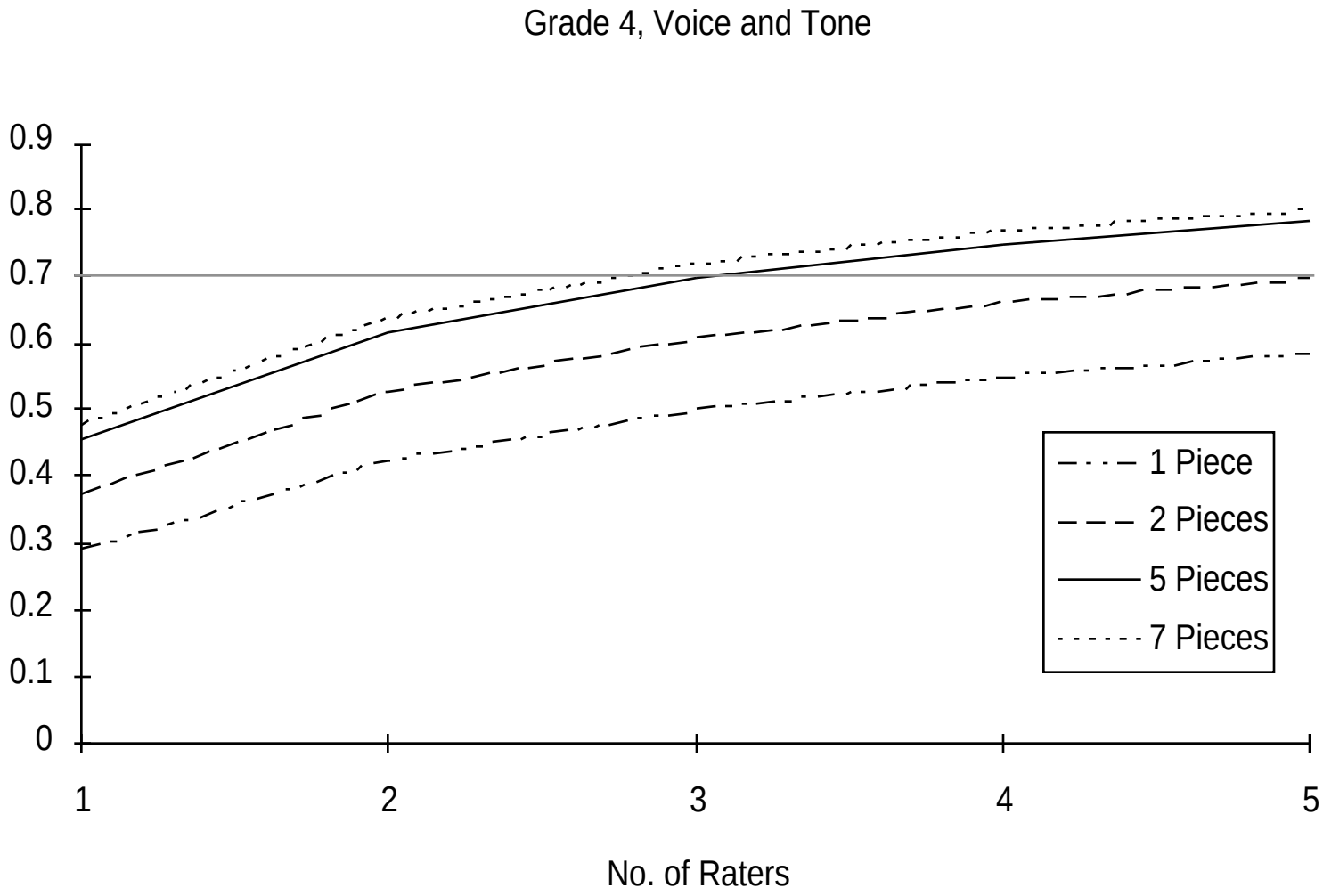


Figure B.6. Effects on Reliability of Increasing the Number of Raters or Pieces,

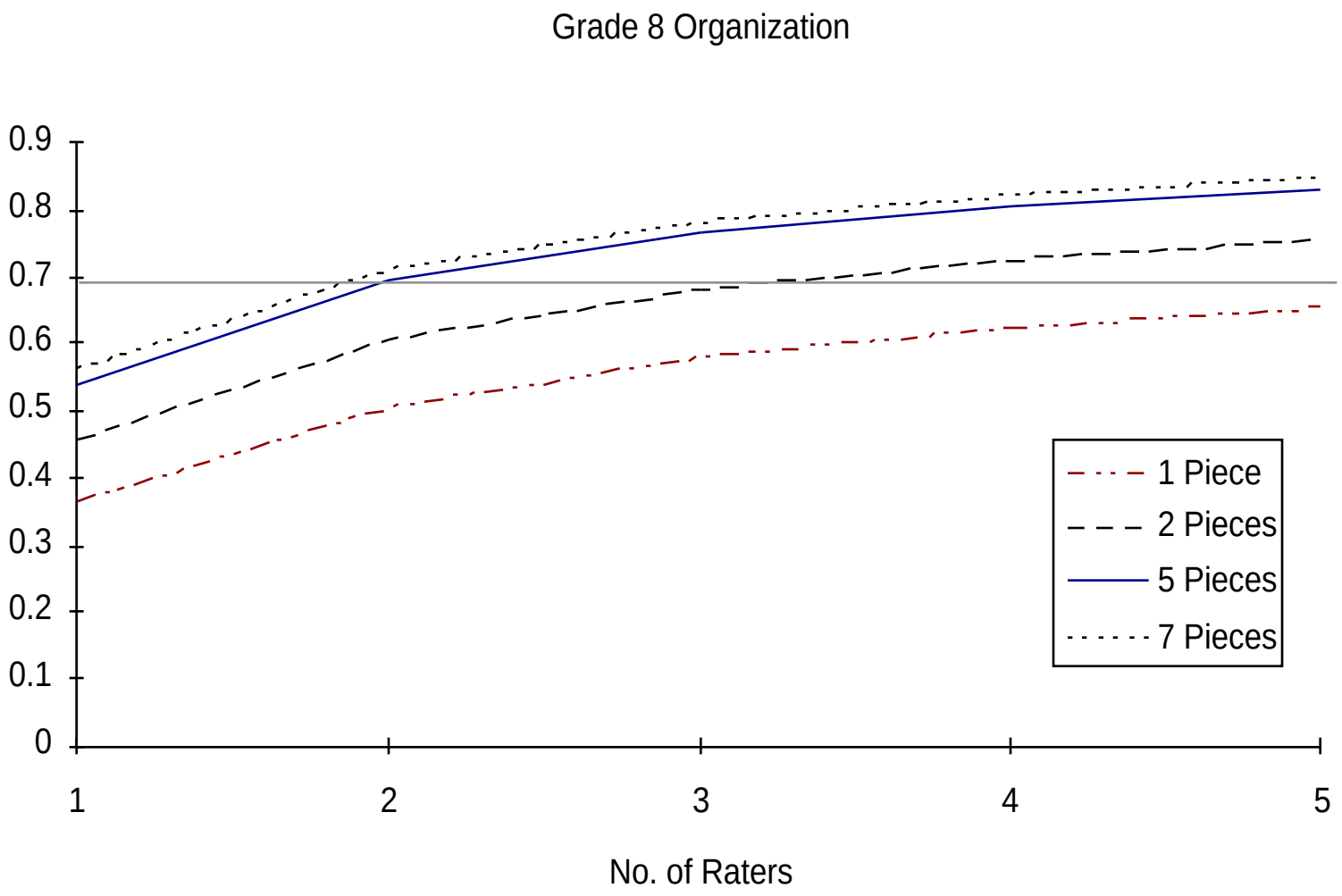


Figure B.7. Effects on Reliability of Increasing the Number of Raters or Pieces,

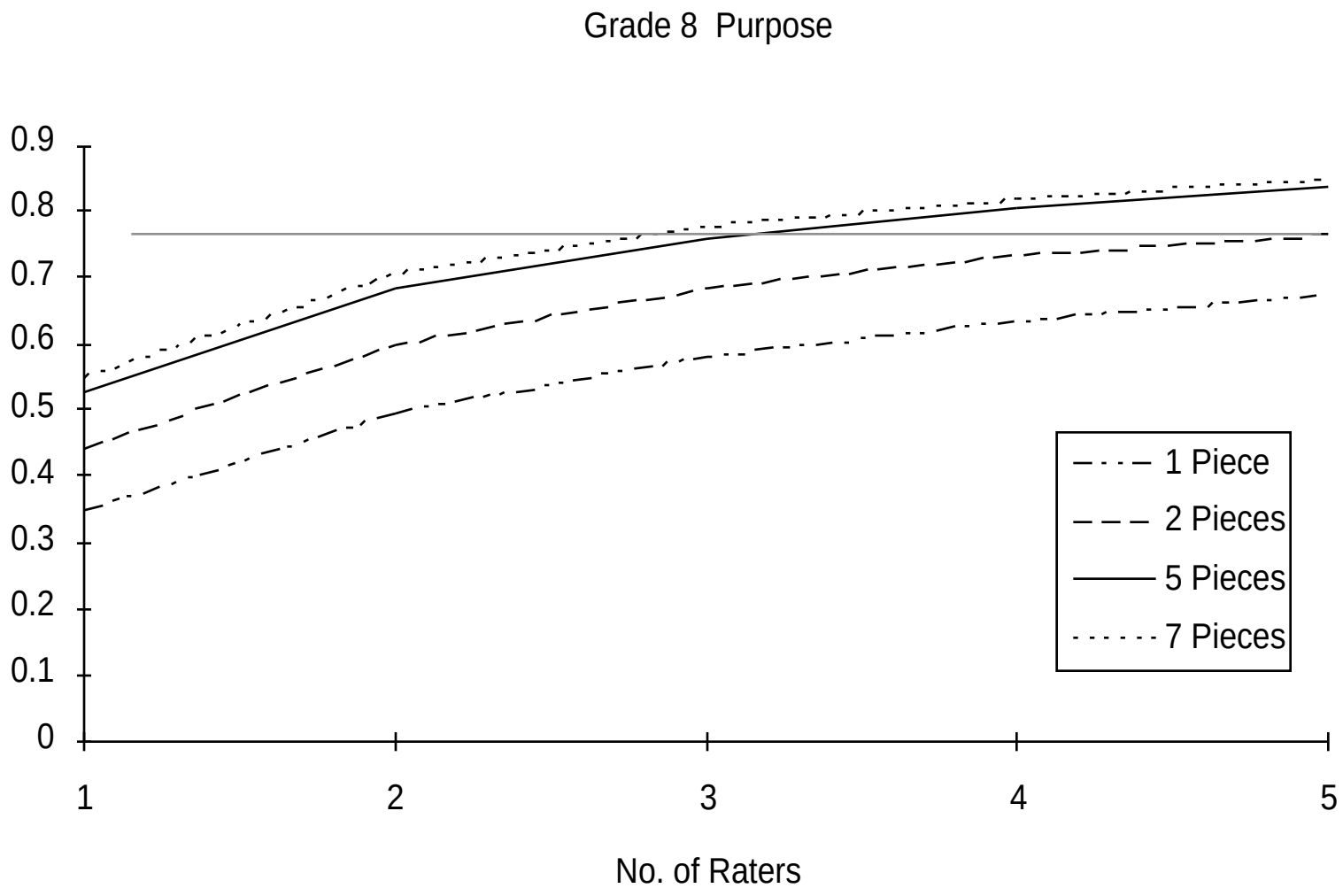


Figure B.8. Effects on Reliability of Increasing the Number of Raters or Pieces,

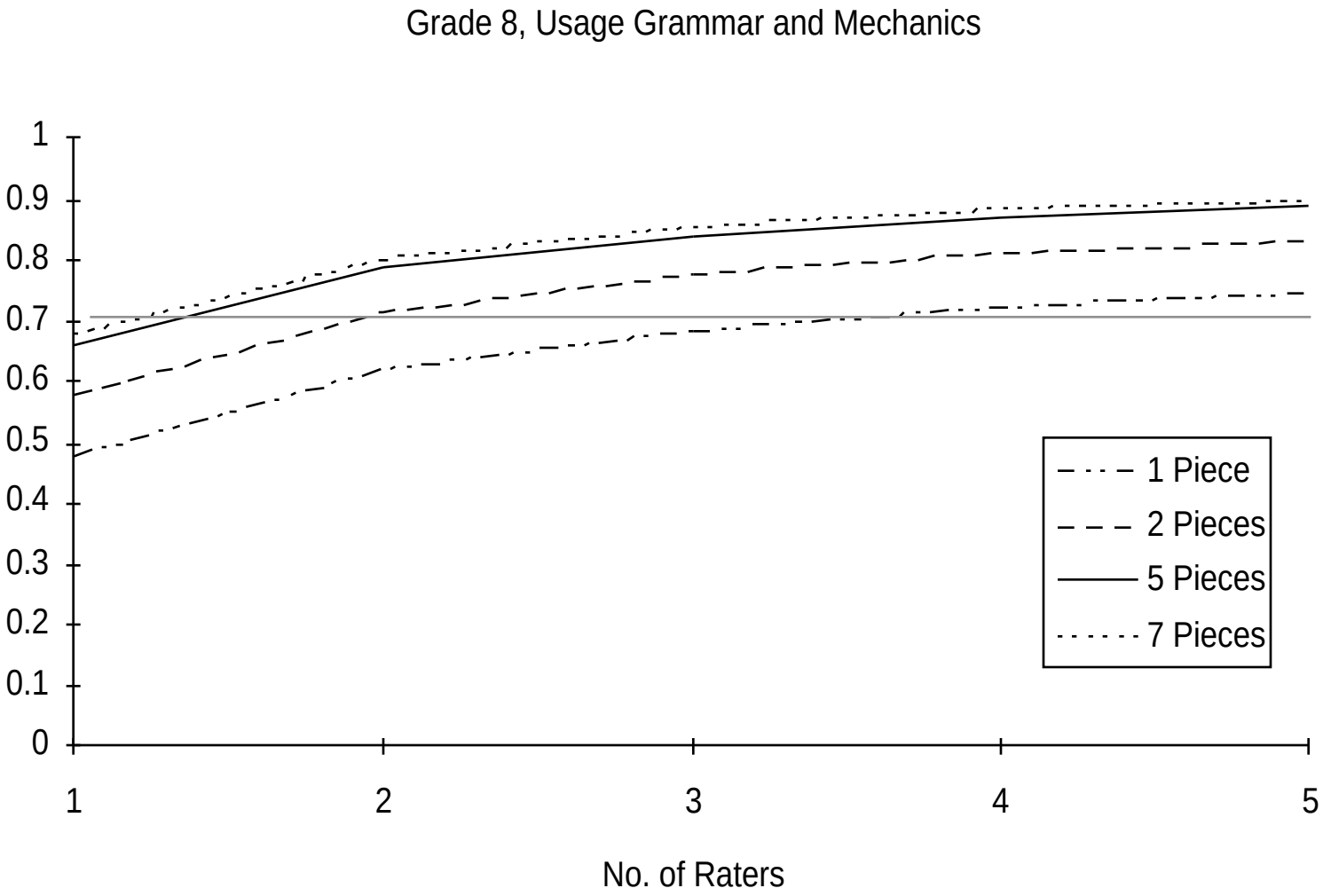


Figure B.9. Effects on Reliability of Increasing the Number of Raters or Pieces,

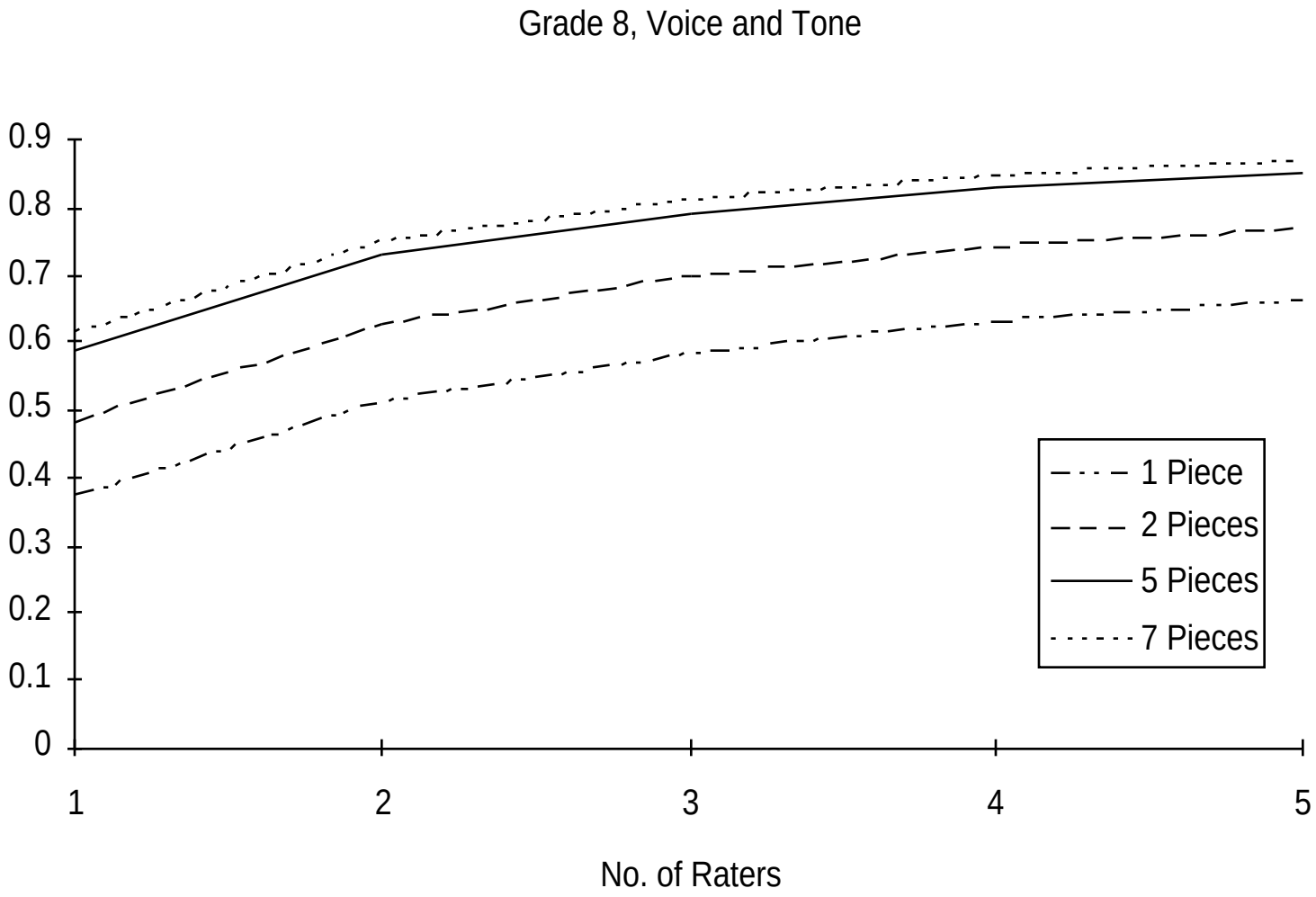
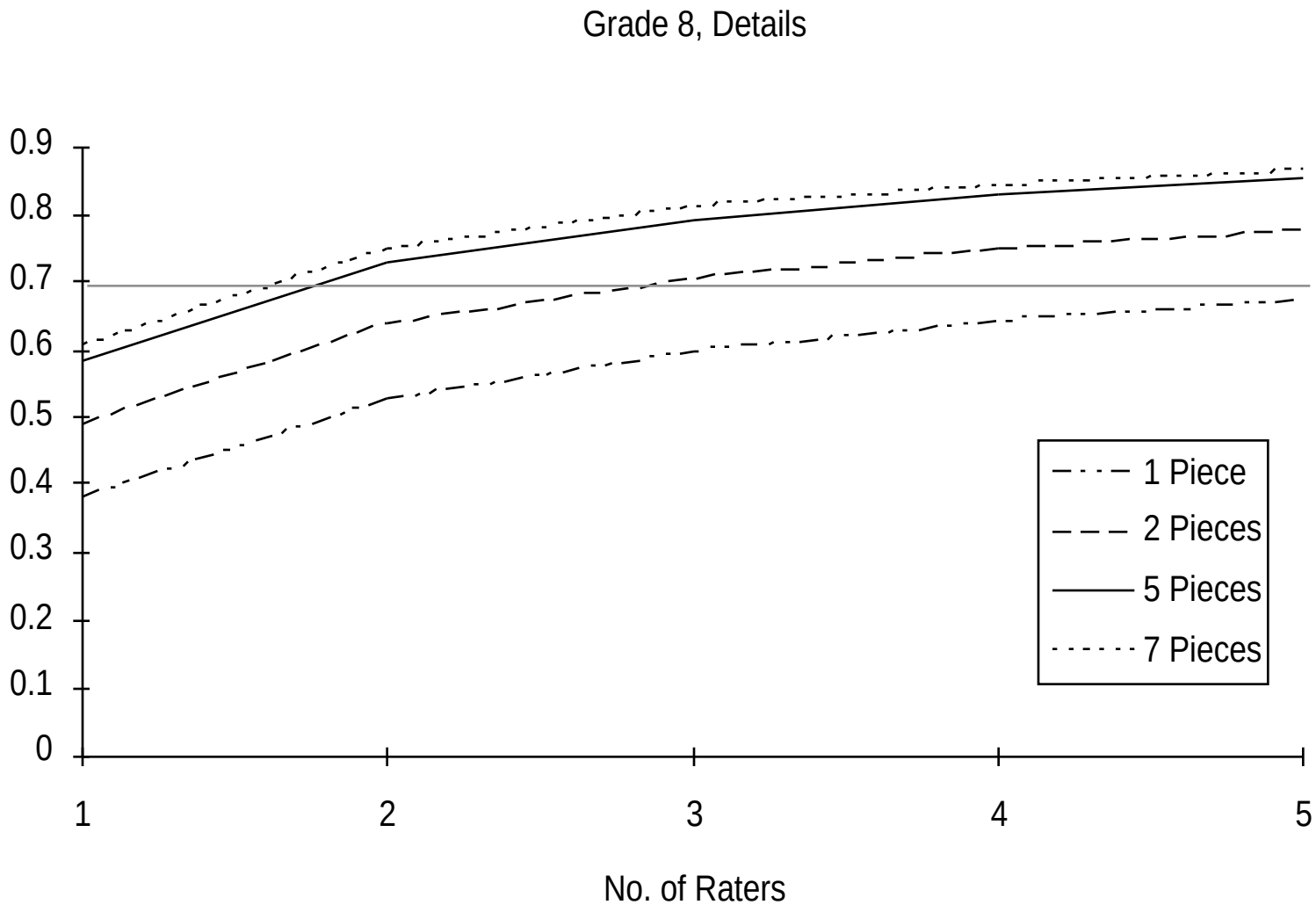


Figure B.10. Effects on Reliability of Increasing the Number of Raters or Pieces,



APPENDIX C

The generalizability study conducted with the math portfolio data is analogous to the study conducted on the writing portfolios. The facets of this study are piece and rater. In other words, the score a student receives on a given portfolio scored by a given rater must generalize to the student's universe score from all portfolios constructed of similar pieces and scored by all similar raters.

In the sample of piece scores each score depends on three factors, the student and the two generalizable facets, rater and piece. Thus the score can be modeled as

$$
\begin{aligned}
y_{ijk} = {} & \mu \\
& + \mu_i - \mu && \text{(person effect)} \\
& + \mu_j - \mu && \text{(rater effect)} \\
& + \mu_{ij} - \mu_i - \mu_j + \mu && \text{(person by rater effect)} \\
& + \mu_{ik} - \mu_i && \text{(piece within person effect)} \\
& + y_{ijk} - \mu_{ij} - \mu_{ik} \\
& \quad + \mu_i + \mu && \text{(residual effect)}
\end{aligned}
$$

where i = 1 to the number of students (803 for Grade 4 and 355 for Grade 8), j = 1 to the number of raters and k = 1 (best piece) or 2 (rest). Only the subset of student portfolios the raters of which could be identified were included in this analysis. The procedure used for scoring student portfolios did not produce a fully crossed data set. Each portfolio was scored by only two raters, not by all raters. This creates an unbalanced data set.

In this study pieces are nested within the student. That is, there is no meaningful interpretation of pieces outside the student's portfolio. In addition, piece number has no meaning. It is an arbitrary naming convention reflecting the order in which students entered pieces in their portfolios. Pieces could be randomly re-assigned different labels without changing the portfolio. This differs from writing, where pieces (or parts) were identified as either a best piece or the "rest" (the remaining pieces, which received a single score). Each piece in the math portfolio received a separate score on each of the seven scoring dimensions. These five to seven scores per dimension constitute a random sample of scores for each student.

However, this is a random sample of scores for pieces the student might select and is probably not a representative sample of scores from all tasks the student might be asked to perform. Students were instructed to select their best work to be included in the portfolio. Thus, if Tiffany is weak at solving probability problems, she may not have included any such problems in her portfolio, and her portfolio scores do not indicate the variability that would result if she had. On the other hand, if Tiffany is weak in solving probability problems, she might never include such a problem in any portfolio she might construct. To the extent that is true, the variability found in these portfolio scores would be representative of the variability that one would expect to find in scores from student-selected portfolios.

Generalizability theory estimates the generalizability of a result by using the observed scores to estimate the variability of each effect or facet. For the math portfolios, we estimated the variability among different raters and the various interaction effects. We also estimated the variability among students. This is considered the true measure of the variation in student ability. This variance is an estimate of the variability that exists in students' "universe" scores—the mean score a student would receive over all possible five- to seven-piece self-selected portfolios scored using the current scoring procedure by all possible similar raters. The interpretation of the universe score, however, must be inferred from other sources such as validity studies.

For each effect given in the above model we estimated the variability of that effect in the population of all such effects. These estimates indicate the amount of the variance in all possible portfolio piece scores by all similar raters that is attributable to each effect. All effects other than the student effect are noise. That is, these effects are the results of the specific portfolio and rater and are not associated with the student's universe score.

Because of the unbalanced nature of the sample of scores, the traditional ANOVA-based estimates of components of variance were not available. Furthermore, the large sample size made it infeasible to use other ANOVA-based methods to estimate the variance components. The component estimates were found using the MIVQUE estimation procedure (Hartley, Rao, & LaMotte, 1978). This method produces unbiased, (locally) minimum variance estimates of the variance components.

The estimates of the variances for the effects given in our model for student scores are in Table C.1 for Grade 4 and Table C.2 for Grade 8. The estimates are found separately for each dimension.

The generalizability of the portfolio score for a single dimension is measured using the generalizability coefficient (Shavelson & Webb, 1991). The generalizability coefficient is approximately equal to the expected value (average) of the square of the correlation between the observed scores and the student's universe scores (Shavelson & Webb, 1991). It is also approximately equal to the correlation between observed scores on two analogous portfolios. For example, the generalizability coefficients given in Tables C.1 and C.2 are estimates of the correlation between two portfolios selected by the same student, each scored by a single rater who reads all five pieces and assigns a score to each piece. This differs from the correlation coefficients given in Chapter 4 because this coefficient measures the degree of agreement between similar portfolios and raters simultaneously, rather than measuring agreement between raters on the same portfolio.

Table C.1

Grade 4 Math Variance Component Estimates

	C1		C2		C3	
Source	**Est.**	**%**	**Est.**	**%**	**Est.**	**%**
Student	**0.0403**	**8.25**	**0.0467**	**7.14**	**0.1480**	**20.87**
Rater	**0.0743**	**15.21**	**0.0498**	**7.62**	**0.0729**	**10.28**
Rater by Student	**0.0313**	**6.40**	**0.0400**	**6.13**	**0.0253**	**3.57**
Piece within Rater	**0.13122**	**27.04**	**0.2348**	**35.92**	**0.1331**	**18.78**
Residual	**0.2106**	**43.09**	**0.2823**	**43.19**	**0.3297**	**46.50**

	PS1		PS2		PS3		PS4	
	Est.	**%**	**Est.**	**%**	**Est.**	**%**	**Est.**	**%**
Student	**0.0505**	**13.06**	**0.0860**	**14.82**	**0.1571**	**20.24**	**0.0211**	**10.82**
Rater	**0.0181**	**4.69**	**0.0228**	**3.93**	**0.1235**	**15.92**	**0.0136**	**6.97**
Rater by Student	**0.0258**	**6.67**	**0.0281**	**4.85**	**0.0213**	**2.74**	**0.0182**	**9.30**
Piece within Student	**0.0747**	**19.32**	**0.1152**	**19.85**	**0.1197**	**15.43**	**0.0489**	**25.01**
Residual	**0.2175**	**56.25**	**0.3282**	**56.55**	**0.3543**	**45.67**	**0.0936**	**47.90**

Table C.2

Grade 8 Math Variance Component Estimates

	C1		C2		C3			
Source	**Est.**	**%**	**Est.**	**%**	**Est.**	**%**		
Student	**0.0583**	**11.7**	**0.0399**	**5.78**	**0.1849**	**22.80**		
Rater	**0.0519**	**10.4**	**0.0523**	**7.58**	**0.0454**	**5.60**		
Rater by Student	**0.0559**	**11.2**	**0.0301**	**4.36**	**0.0708**	**8.73**		
Piece within Rater	**0.1197**	**24.1**	**0.2827**	**40.99**	**0.1810**	**22.32**		
Residual	**0.2115**	**42.5**	**0.2847**	**41.28**	**0.3289**	**40.56**		
	PS1		**PS2**		**PS3**		**PS4**	
	Est.	**%**	**Est.**	**%**	**Est.**	**%**	**Est.**	**%**
Student	**0.0435**	**11.59**	**0.0526**	**9.70**	**0.1117**	**14.1**	**0.0153**	**10.59**
Rater	**0.0103**	**2.75**	**0.0236**	**4.35**	**0.0912**	**11.5**	**0.0071**	**4.90**
Rater by Student	**0.0363**	**9.69**	**0.0469**	**8.64**	**0.0829**	**10.5**	**0.0019**	**1.32**
Piece within Student	**0.0884**	**23.59**	**0.1544**	**28.48**	**0.1858**	**23.4**	**0.0345**	**23.83**
Residual	**0.1965**	**52.38**	**0.2648**	**48.83**	**0.3213**	**40.5**	**0.0859**	**59.36**

The advantage of this generalizability analysis is that because we have estimated the components of variance, we can estimate the expected square of the correlation between universe scores and observed scores for various portfolio plans and scoring schemes. For example, we can estimate this correlation for a portfolio with five pieces, scored by two raters, where the score is the average of these ten scores. The score from a portfolio is assumed to be the total (or average score) over all parts and raters for each dimension.

These generalizability coefficients are meaningful only if we assume that the scoring procedure remains constant. That is, each rater scores all pieces together; the independent raters do not each score the parts separately or independently. We must assume the same scoring procedure because changing the scoring procedure might change the variability among scores. Also, this generalizability coefficient is only meaningful if we are considering similar selected pieces. Under these assumptions, we can estimate the generalizability coefficient for any combination of parts and raters. The generalizability coefficients are calculated using the following formula

$$\frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_{sr}^2 / n_r + \hat{\sigma}_{sp}^2 / n_p + \hat{\sigma}_{srp.e}^2 / n_r n_p}$$

where $\hat{\sigma}_i^2$ i = s, sr, sp and srp.e denote the estimated variance components for students, student by rater, piece within student and residual effects respectively, and n_r and n_p denote the number raters scoring each portfolio and the number of pieces respectively.

For Grade 4, Figures C.1 to C.7 show the effects on the coefficient from varying the number of parts and the number of raters. Figures C.8 to C.14 are the analogous plots for Grade 8.

Figure C.1. Effects on Reliability of Increasing the Number of Raters or Pieces,

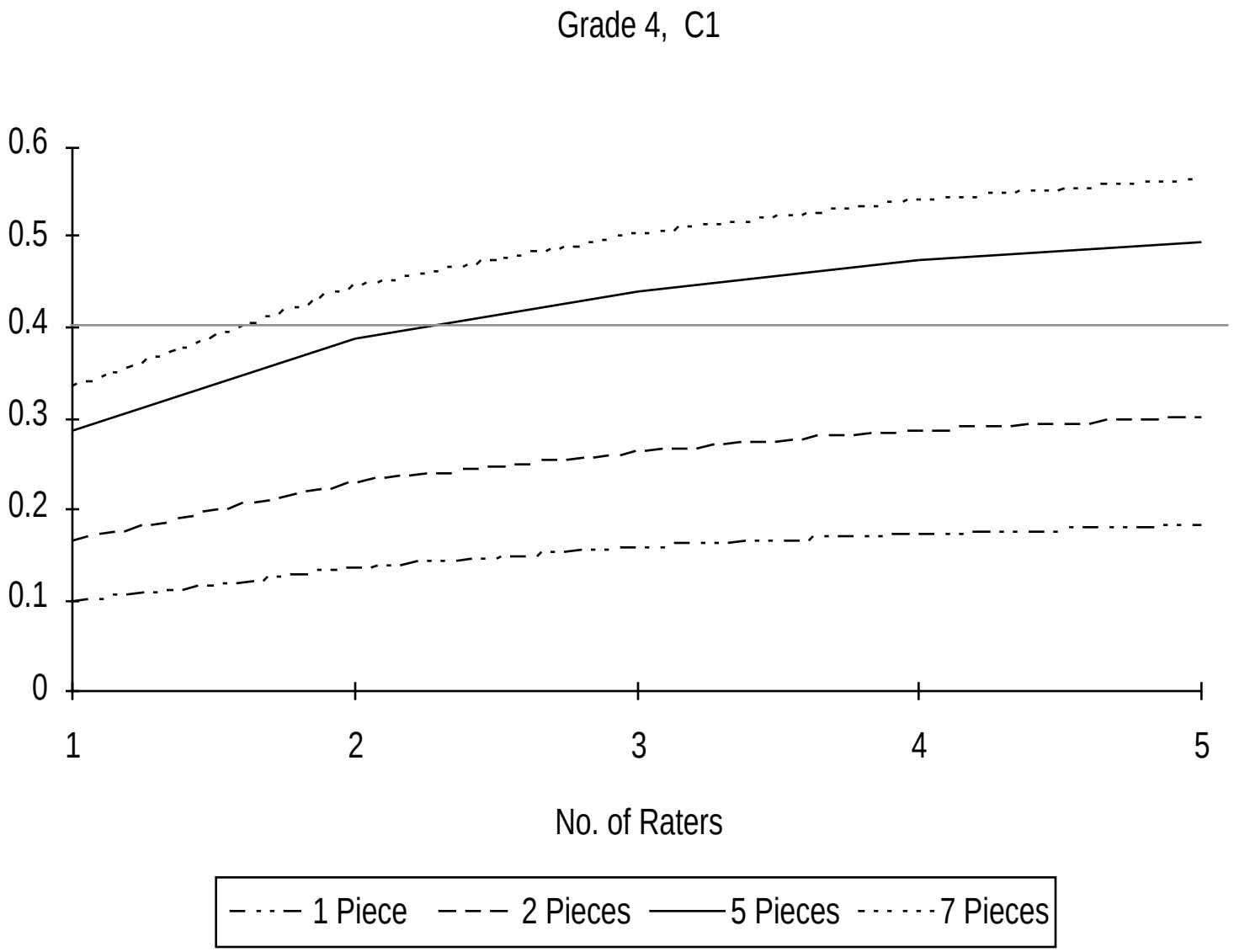


Figure C.2. Effects on Reliability of Increasing the Number of Raters or Pieces,

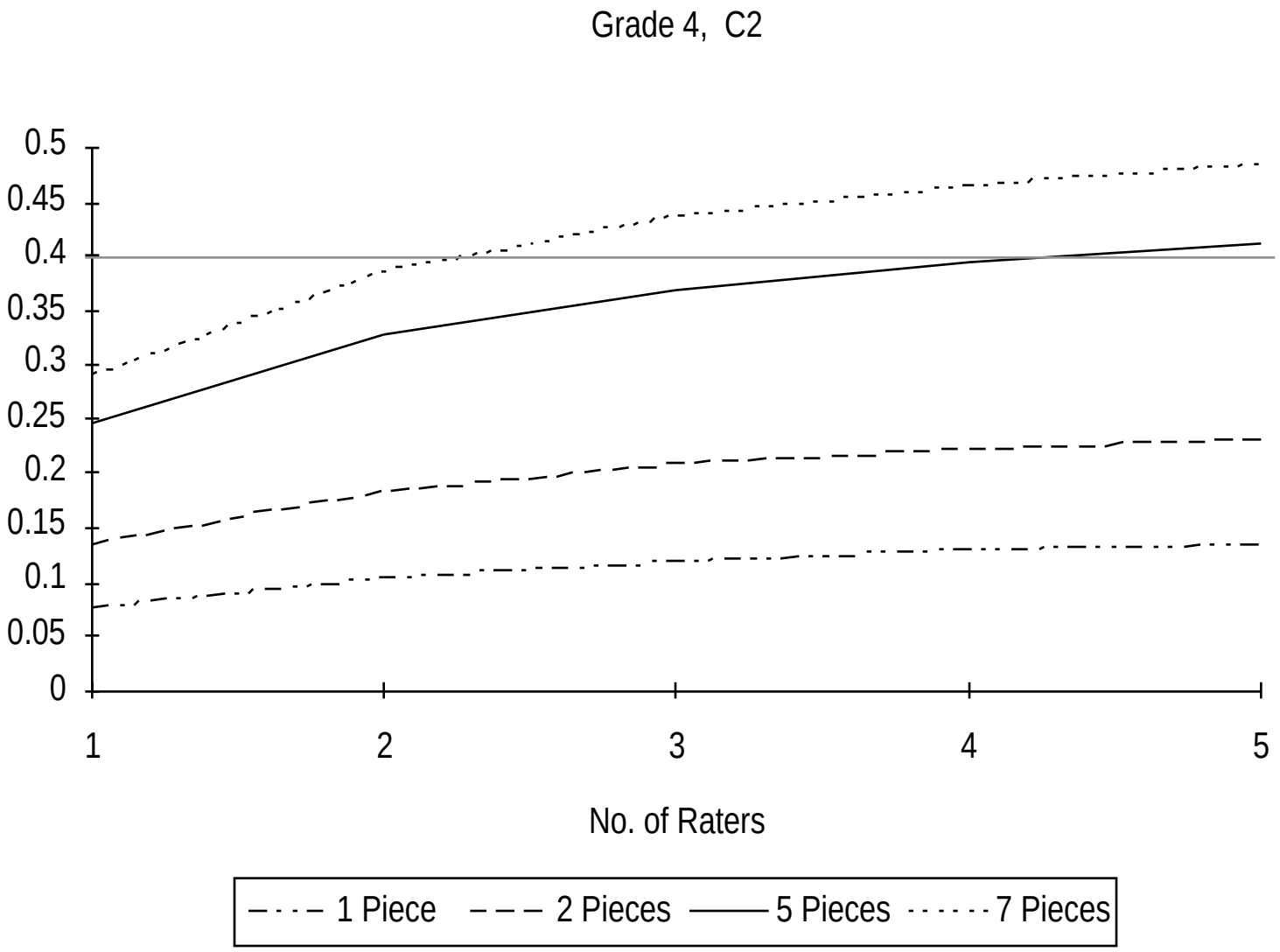


Figure C.3. Effects on Reliability of Increasing the Number of Raters or Pieces,

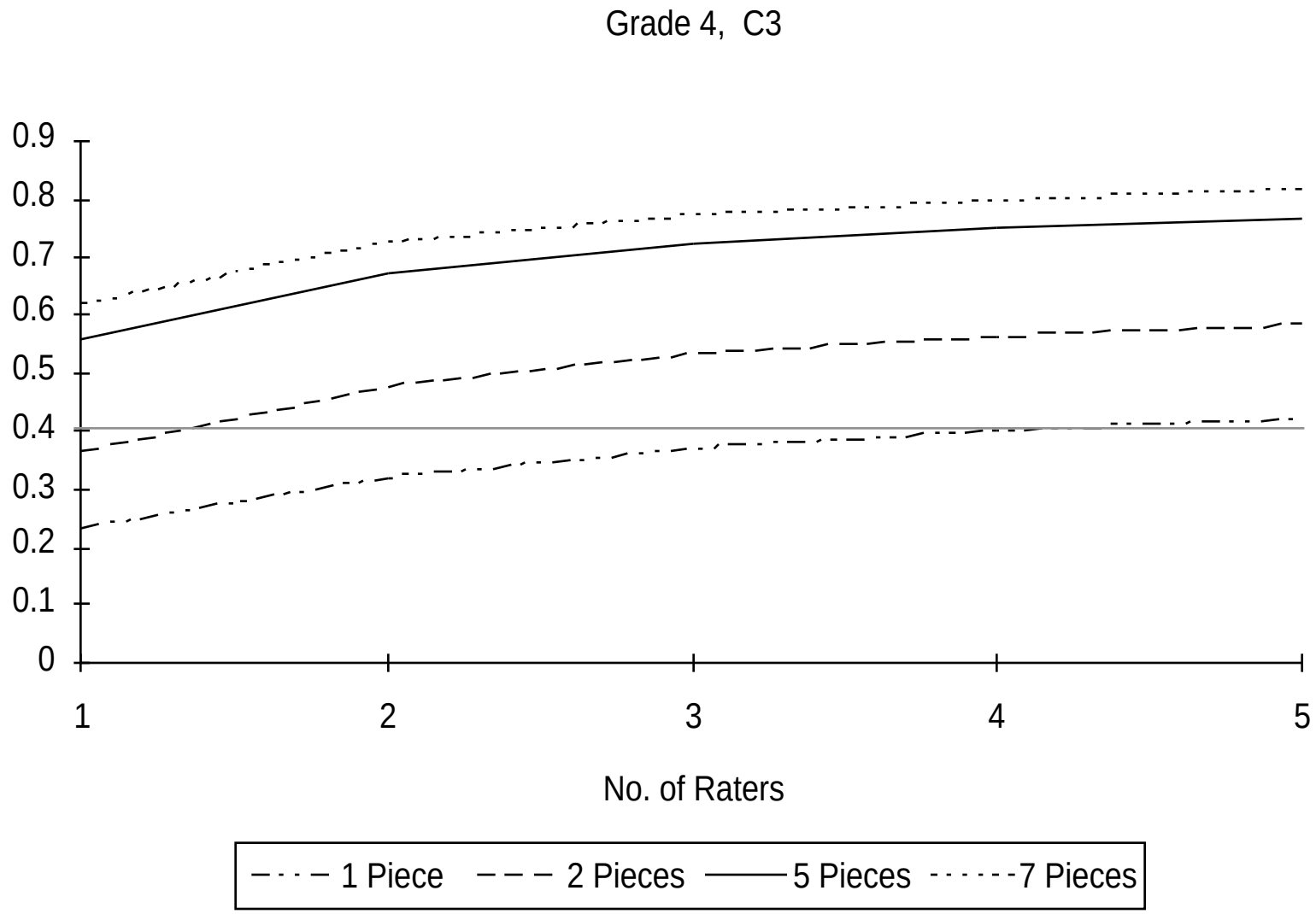


Figure C.4. Effects on Reliability of Increasing the Number of Raters or Pieces,

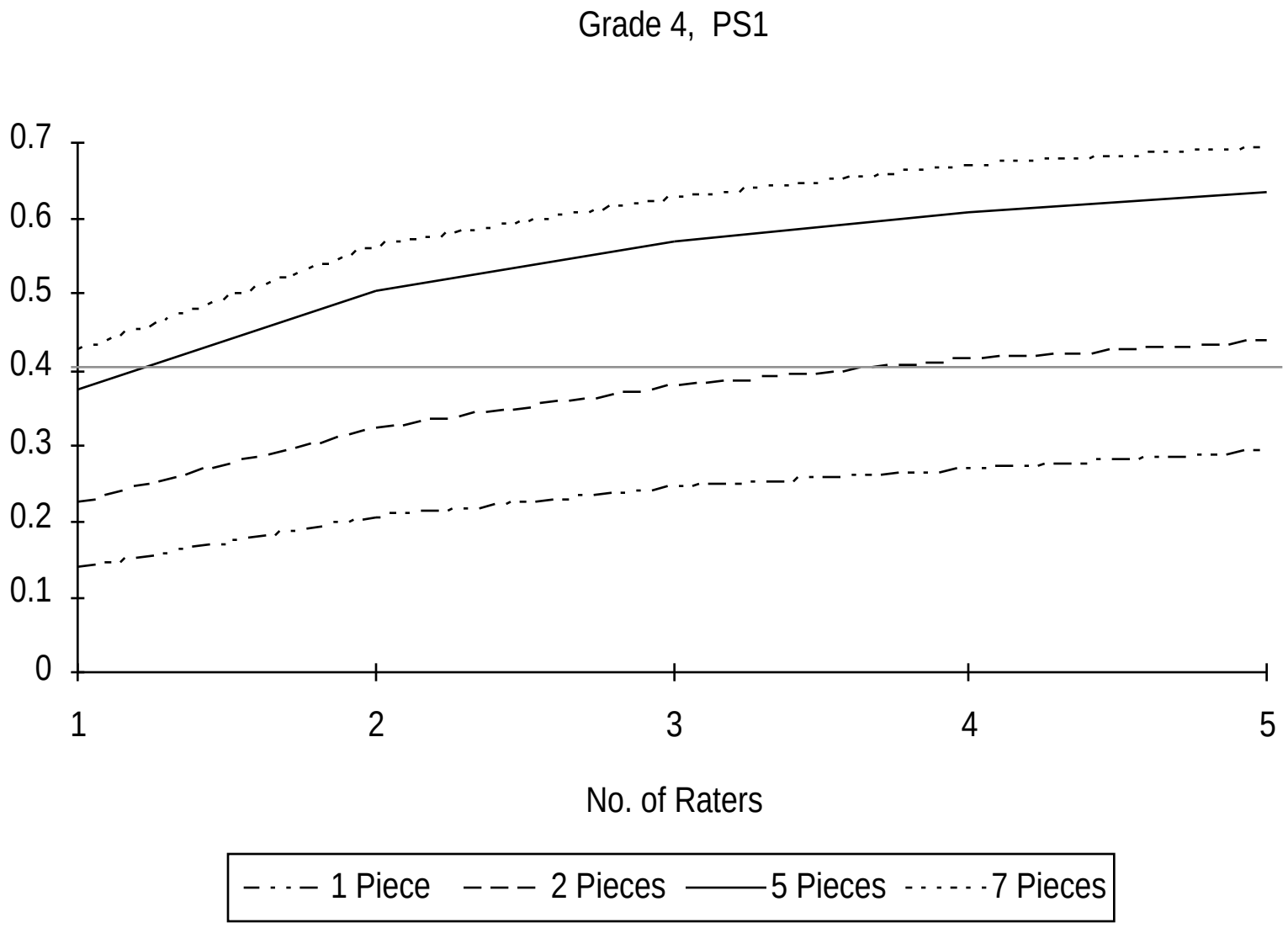


Figure C.5. Effects on Reliability of Increasing the Number of Raters or Pieces,

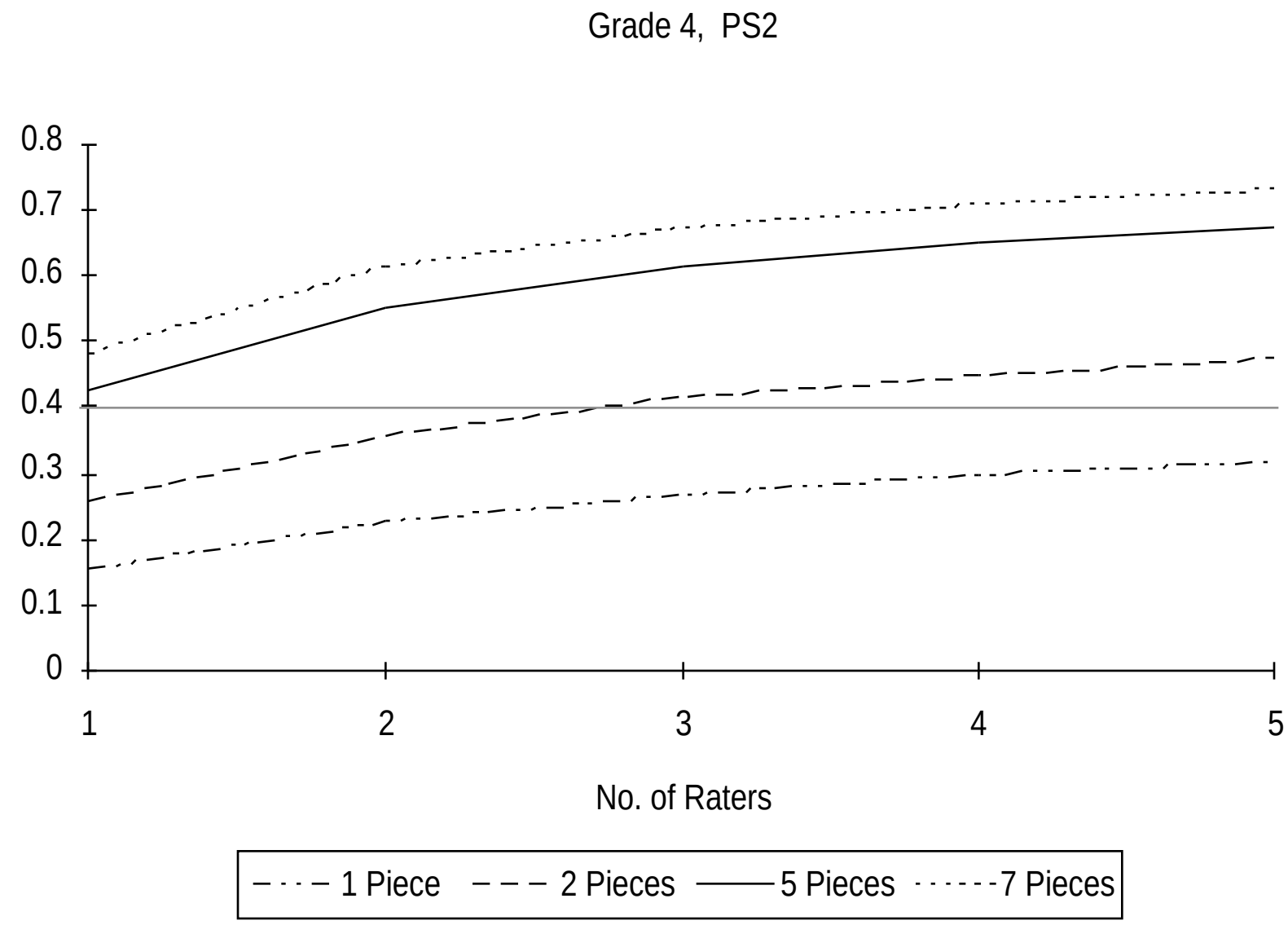


Figure C.6. Effects on Reliability of Increasing the Number of Raters or Pieces,

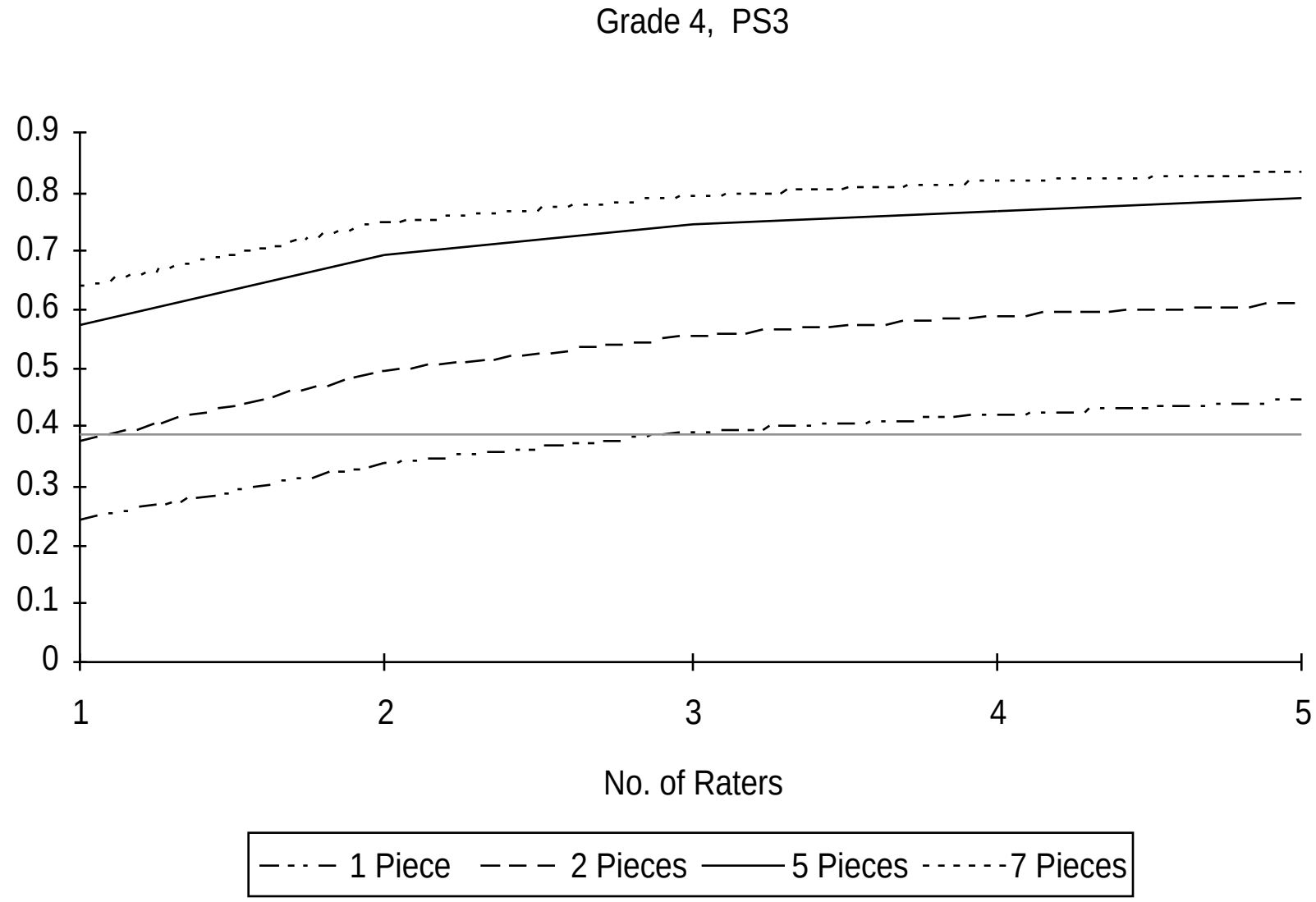


Figure C.7. Effects on Reliability of Increasing the Number of Raters or Pieces,

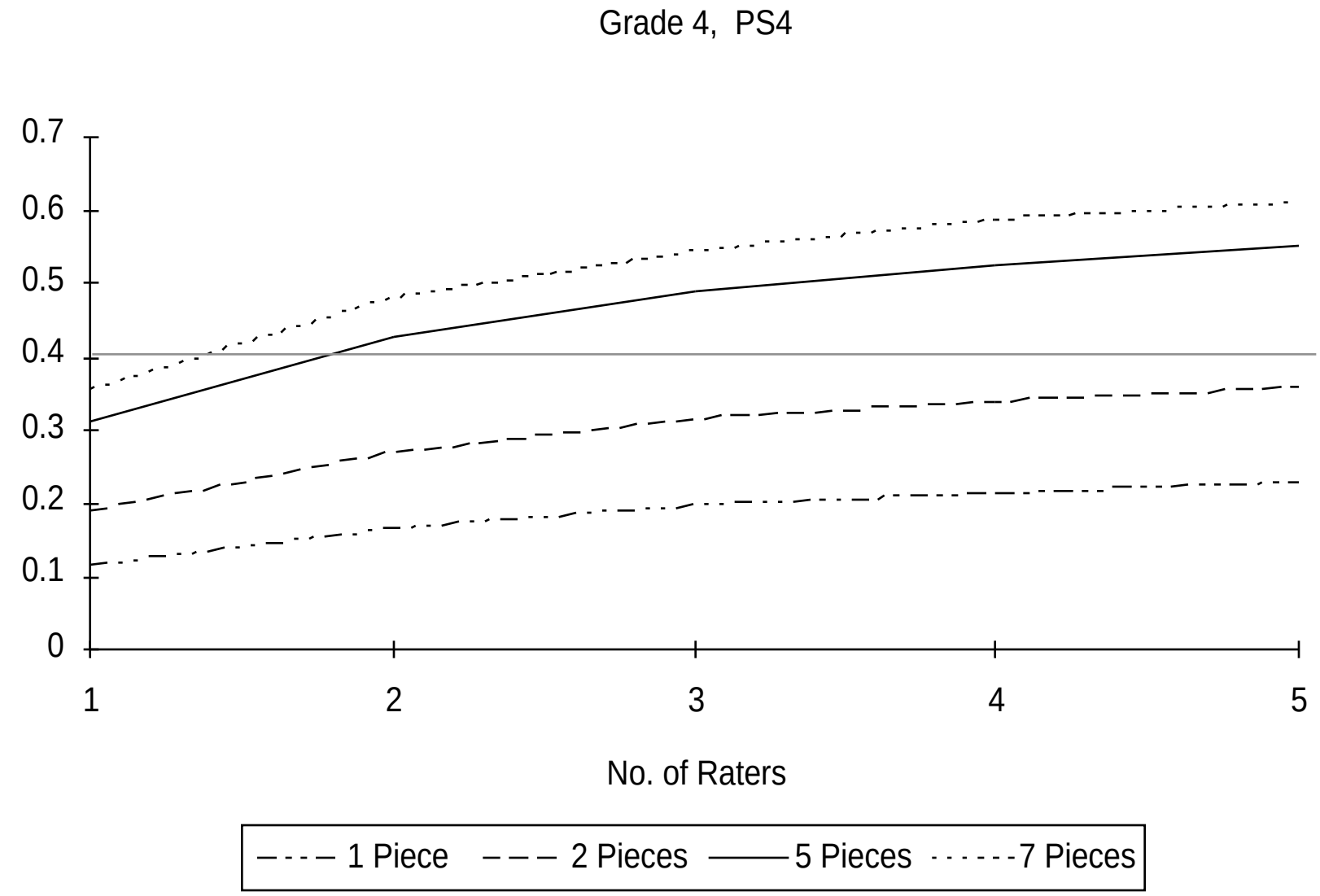


Figure C.8. Effects on Reliability of Increasing the Number of Raters or Pieces,

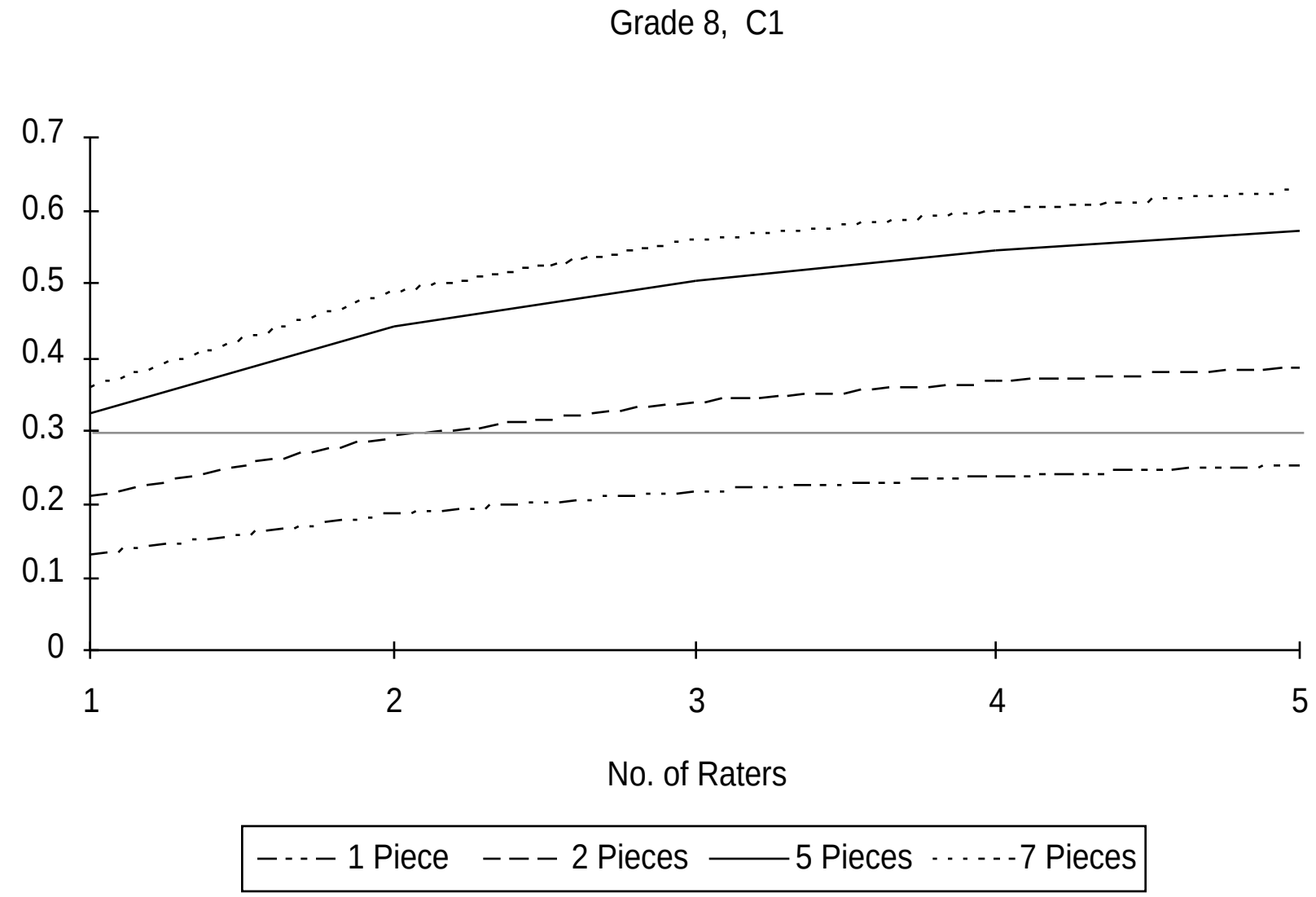


Figure C.9. Effects on Reliability of Increasing the Number of Raters or Pieces,

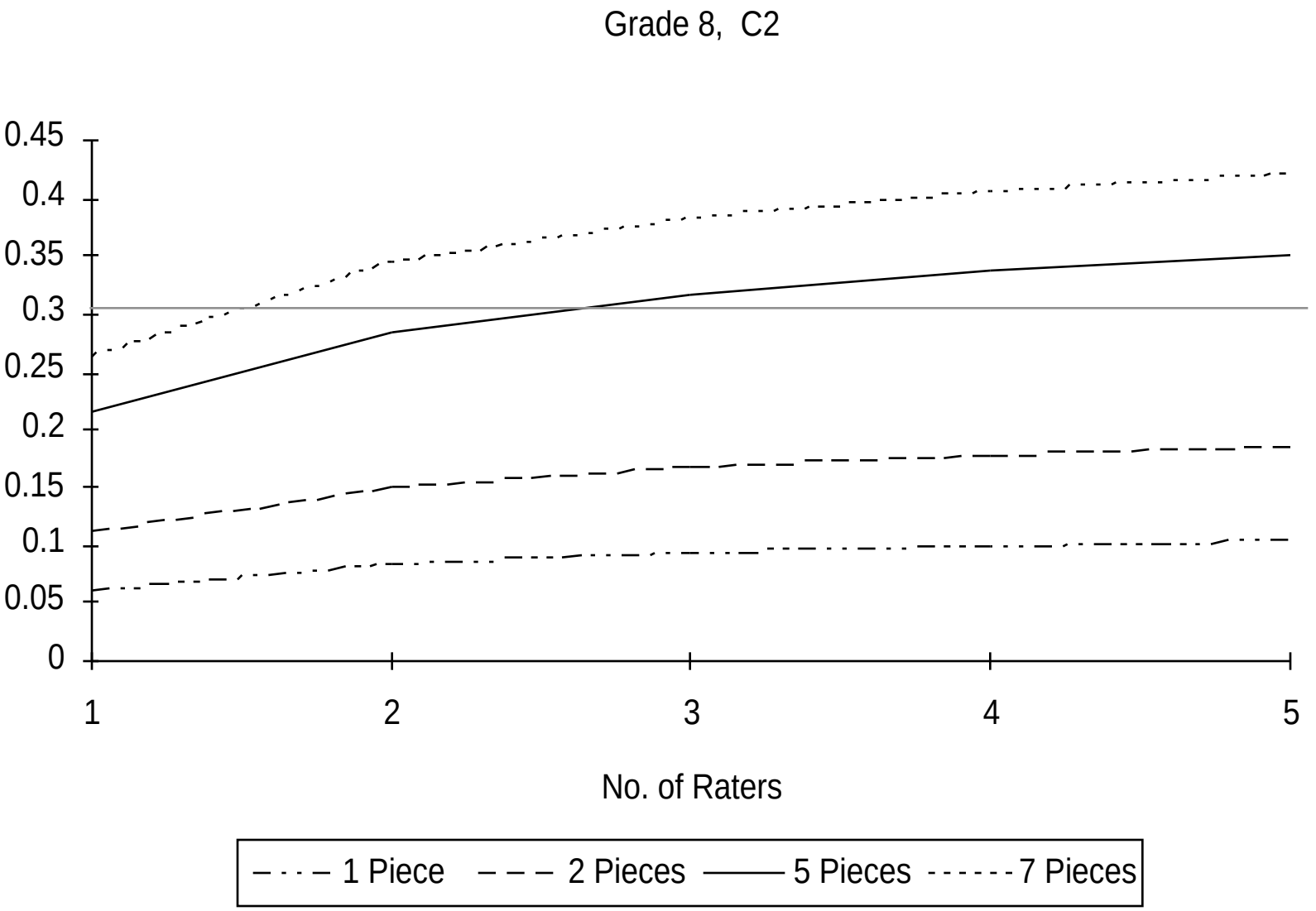


Figure C.10. Effects on Reliability of Increasing the Number of Raters or Pieces,

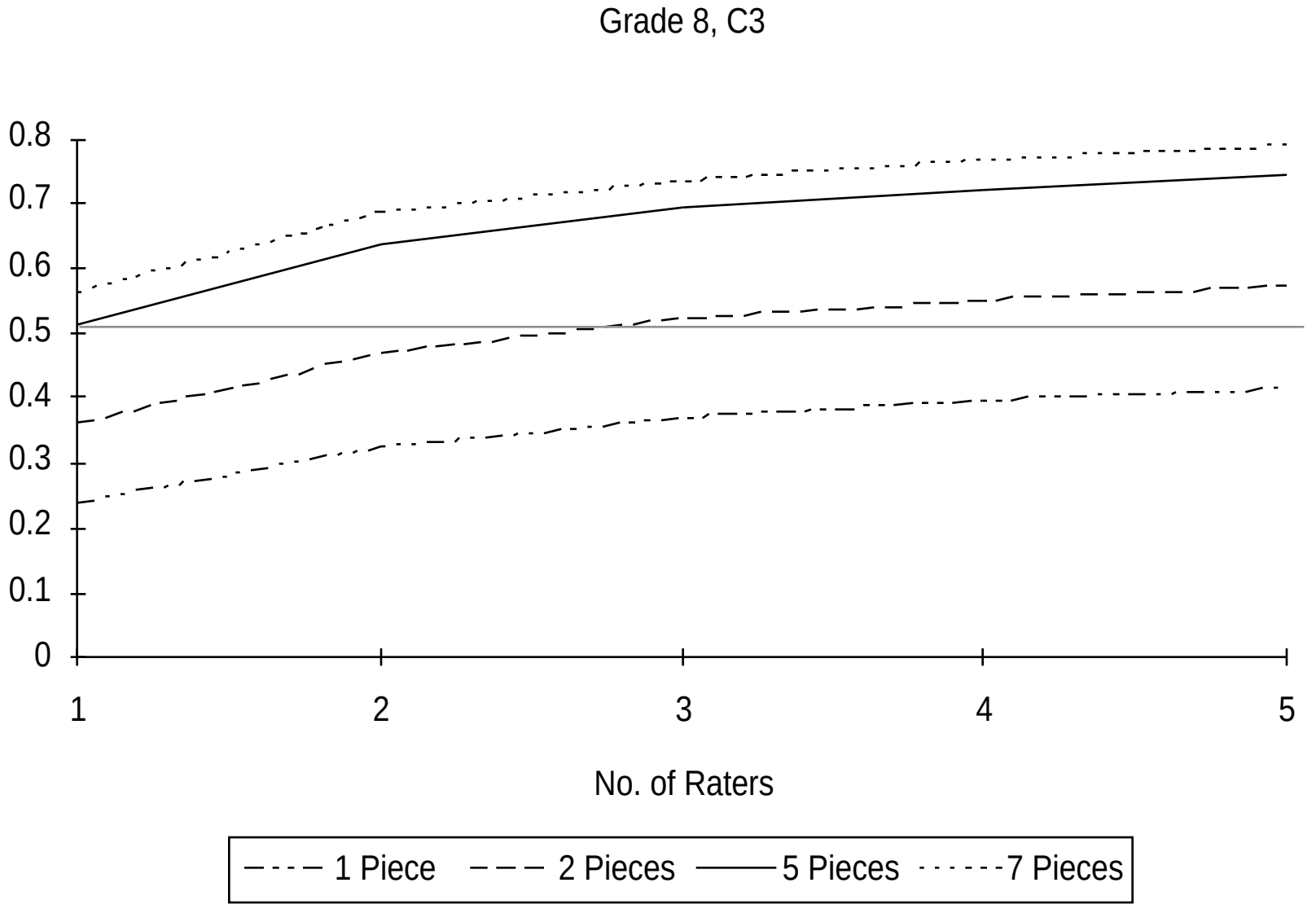


Figure C.11. Effects on Reliability of Increasing the Number of Raters or Pieces,

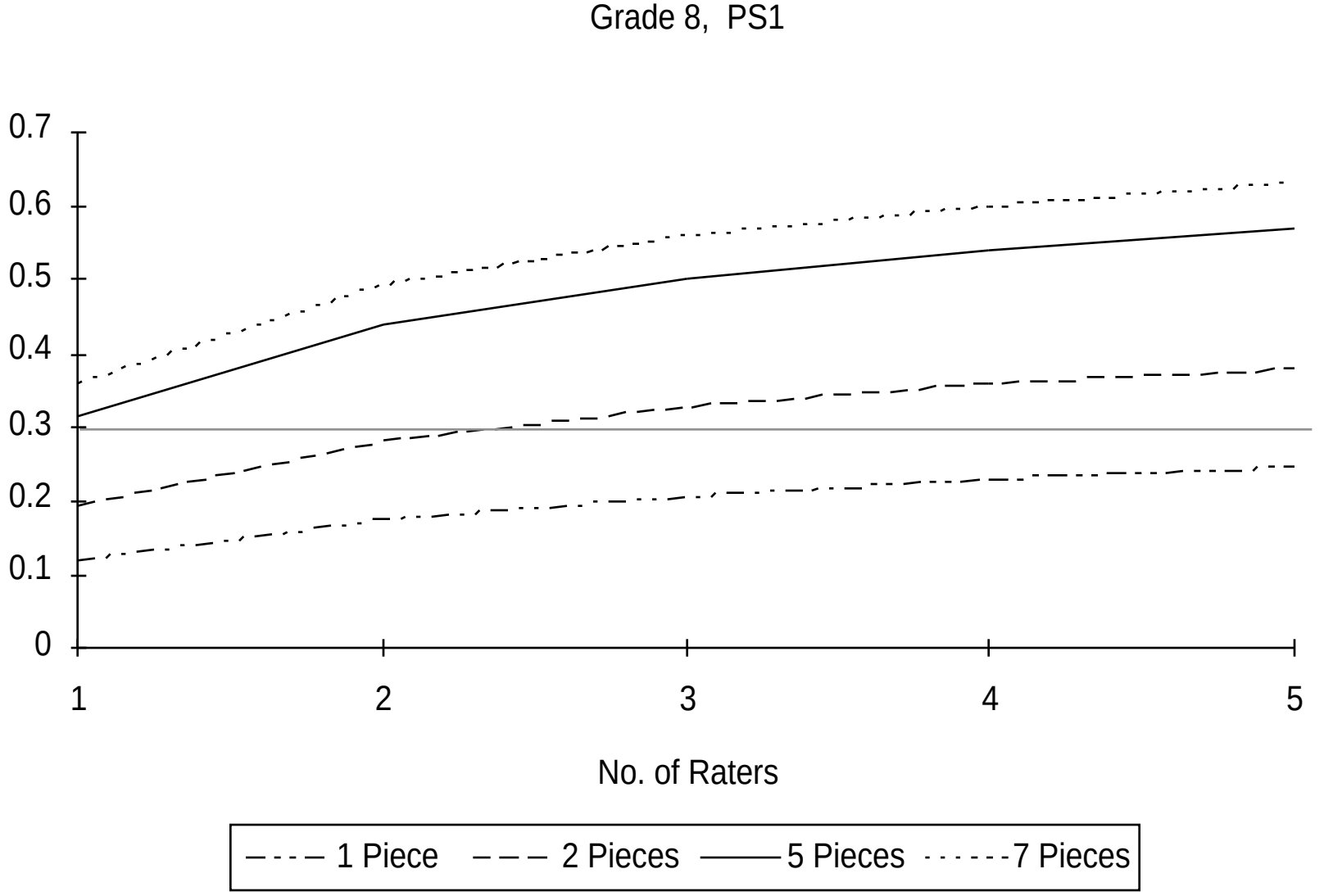


Figure C.12. Effects on Reliability of Increasing the Number of Raters or Pieces,

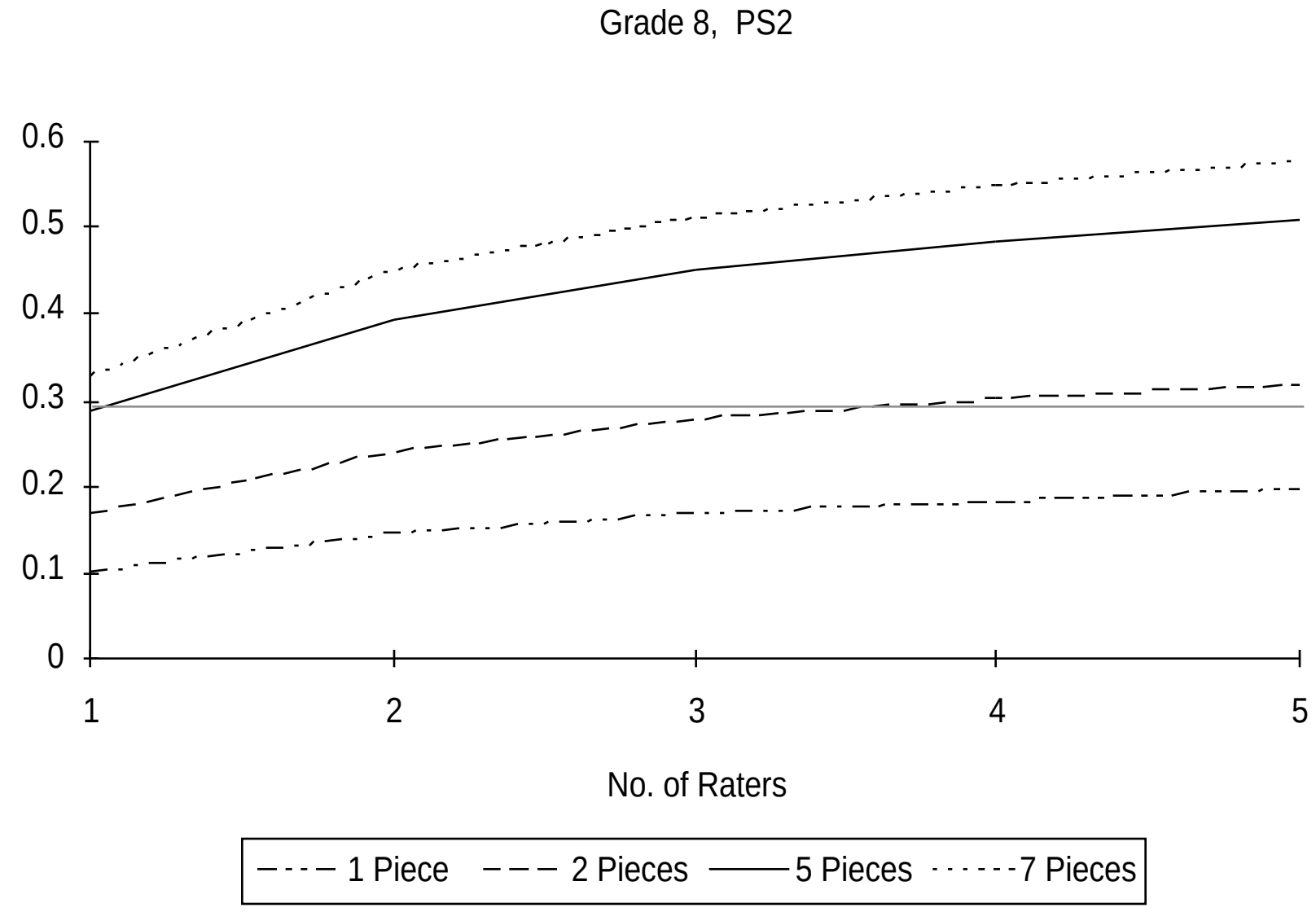


Figure C.13. Effects on Reliability of Increasing the Number of Raters or Pieces,

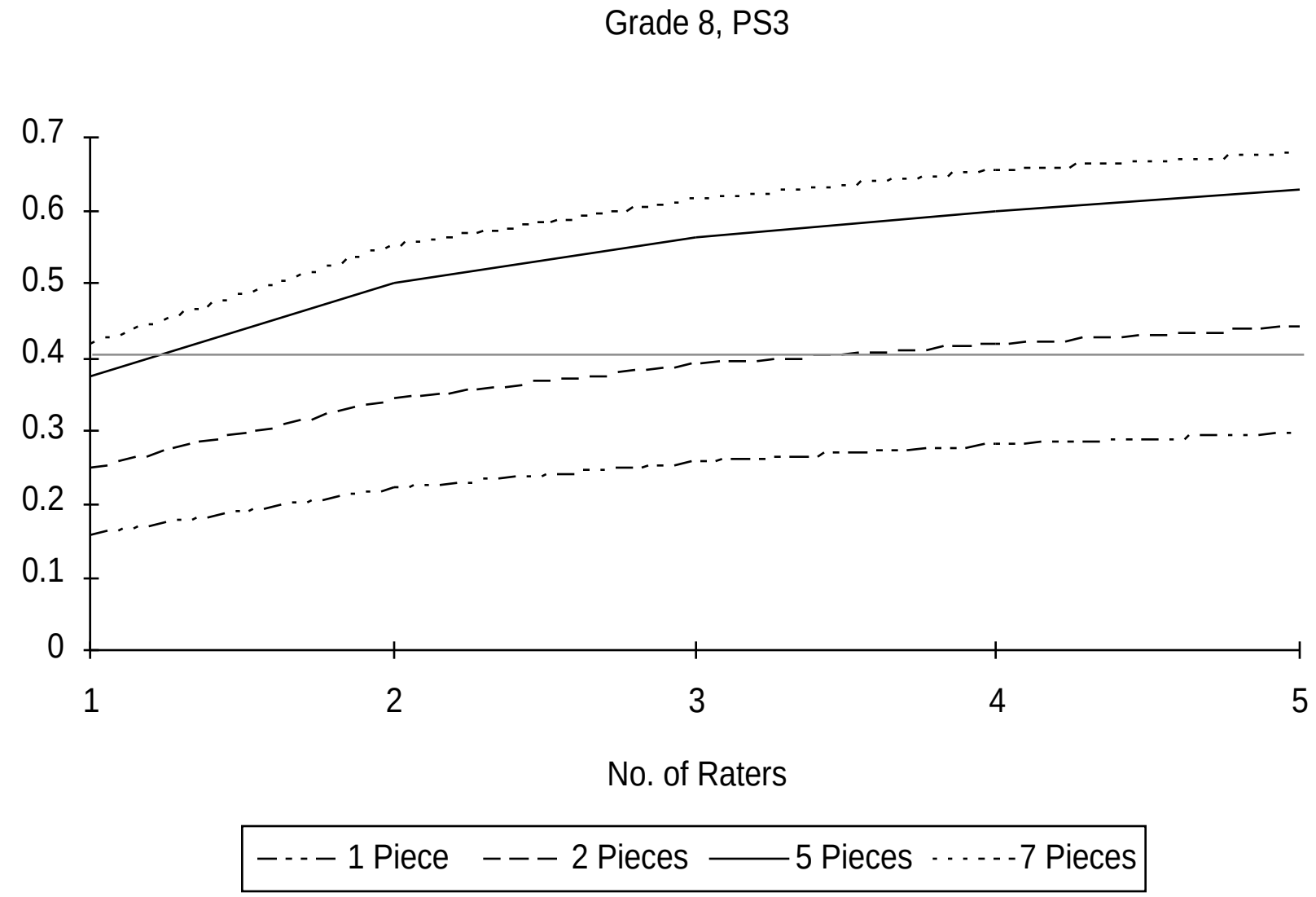
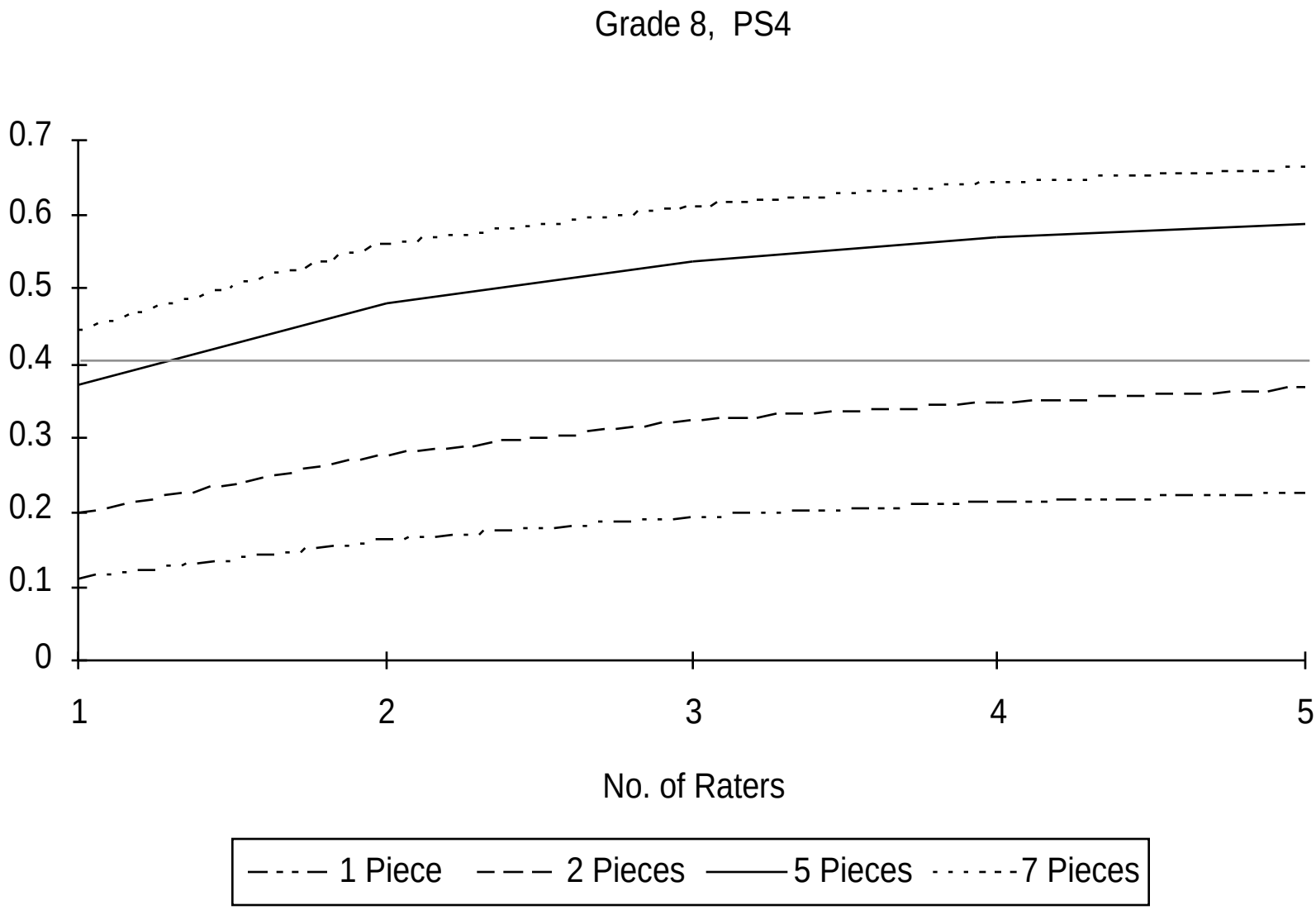


Figure C.14. Effects on Reliability of Increasing the Number of Raters or Pieces,



APPENDIX D. SCORING WORKSHEETS