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

Generalizability of New Standards Project 1993 Pilot Study Tasks in Mathematics

CSE Technical Report 392

Robert L. Linn and Elizabeth Burton
CRESST/University of Colorado at Boulder

Lizanne DeStefano and Matthew Hanson
University of Illinois at Urbana-Champaign

January 1995

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1521
(310) 206-1532

The work reported herein was partially 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. It was also conducted with partial support from the John D. and Catherine T. MacArthur Foundation and the Pew Charitable Trusts through the New Standards Project. The findings and opinions expressed in this report do not reflect the policies of the Office of Education Research and Improvement or the U.S. Department of Education or of the other aforementioned funding agencies.

GENERALIZABILITY OF NEW STANDARDS PROJECT 1993 PILOT STUDY TASKS IN MATHEMATICS[1]

Robert L. Linn and Elizabeth Burton
CRESST/University of Colorado at Boulder

Lizanne DeStefano and Matthew Hanson
University of Illinois at Urbana-Champaign

Abstract

The New Standards Project conducted a pilot test of a series of performance-based assessment tasks in mathematics and English language arts at Grades 4 and 8 in the spring of 1993. This paper reports the results of a series of generalizability analyses conducted for a subset of the 1993 pilot study data in mathematics. Generalizability analyses for completely crossed designs of raters-by-tasks-by-pupils were conducted for a total of nine collections of mathematics tasks. The results of those analyses were used to estimate generalizability coefficients and standard errors of measurement for decision studies using various combinations of number of raters and number of tasks. Consistent with results of previous analyses of performance-based assessment tasks, sampling variability due to tasks was found to be substantially larger than that due to raters. Implications for assessment designs are discussed.

Performance-based assessments have a prominent role in contemporary educational reform efforts. The use of performance assessments as a reform tool is based, in part, on the argument that improvements in education can be facilitated by the construction and use of assessments that are worth teaching to (see, for example, Resnick & Resnick, 1992). Although the performance assessment movement has considerable conceptual appeal, it also poses a number of questions regarding the technical quality of results. The purpose of this paper is to present results bearing on one of the major technical quality

[1] We thank members of the New Standards Project Technical Design Committee (Dale Carlson, Ruben Carriedo, Lee Cronbach [through March, 1994], Richard Durán, Andrew Porter, and Floraline Stephens) for their contributions to the study design and helpful comments on an earlier draft of this report.

issues, namely, the generalizability of results as a function of raters and tasks, for assessments of mathematics developed and piloted by the New Standards Project.

The New Standards Project

The New Standards Project (NSP) is a joint program of the National Center on Education and the Economy based in Rochester, NY, and the Learning Research and Development Center at the University of Pittsburgh and is co-directed by Mark Tucker and Lauren Resnick. Funded by the Pew Charitable Trusts and the John D. and Catherine T. MacArthur Foundation, the project includes 18 state and 6 district partners. Its partners enroll nearly half of the public school children in the United States.

The NSP endorses a reform strategy based on a system of standards, assessment, professional development, technical assistance, and quality control. Its goal is to improve learning significantly for all students, particularly those who do not perform well now. It includes the areas of mathematics, literacy, science, and applied learning.

The NSP is creating an assessment system that includes an on-demand reference exam and portfolios. The reference exam is intended to provide partners with a performance-based examination reflecting common content and performance standards. Results of student performance will be reported for the New Standards partnership as a whole and for individual partners, enabling partners to gauge student performance to common standards set by the Governing Board. More detailed information on the NSP can be obtained by contacting the National Center on Education and the Economy, 700 Eleventh Street, NW, Suite 750, Washington, DC 20001.

The 1993 Spring Pilot in Mathematics

During April and May, 1993, 28 mathematics tasks (12 at 4th grade and 16 at 8th grade) were piloted with 20,393 students in 880 classrooms in 23 partner states and districts. Although the pilot sample was not selected in ways that permit generalization to any larger group, the students who participated in the 1993 pilot in mathematics were diverse in terms of race/ethnicity, primary language, economic status, and academic performance (DeStefano, 1993b).

The time required for administration of the pilot study mathematics tasks varied from task to task. "Long tasks" required between one and three hours each to administer. On long tasks, students were asked to give complex, multifaceted responses, which often included drawing a graph or figure or writing a paragraph. All of the long tasks included nonscored, pre-assessment activities in the total administration time. Some one-hour blocks of administration time were also used for "short forms" that typically included four shorter tasks that were all to be completed within one hour. The complete set of pilot test mathematics tasks consisted of a total of 11 long tasks and 4 short forms at Grade 4 and 16 long tasks and 4 short forms at Grade 8. Long tasks were scored on a 4-point scale. Short tasks were scored dichotomously (correct/incorrect). Because a short form consisted of four short tasks, the sum of the scores on the tasks on a short form had a maximum score of 4, and the sum was used to summarize performance on the short forms.

Student responses were scored during the Summer Leadership Conference held July 7-14 in Snowbird, Utah. Approximately 150 participants were trained and scored the mathematics tasks that were administered during the spring pilot. Most scorers were teachers or curriculum supervisors. They were highly experienced in terms of classroom teaching (mode = more than 20 years of teaching), but more than half had never scored students' written responses to open-ended questions or prompts in a large-scale, standardized manner. Ninety-four percent of the scorers qualified by scoring 16 out of 20 student responses consistently with the benchmark set (DeStefano, 1993b). Only scores from qualified scorers were considered in this study.

The interrater agreement results for the ratings of New Standards Project (NSP) tasks administered in the 1993 pilot test were summarized by DeStefano (1993a, 1993b). With the exception of one task, the percent exact agreement between raters on the 4-point scales used for the long mathematics tasks ranged between 60% and 75%.

Although raters are one important source of measurement error, experience with other performance-based assessments suggests that variability due to the sampling of tasks is usually greater than that due to raters. Thus, it is important to evaluate the degree of generalizability of scores across both raters and tasks. This report presents the results of the generalizability analyses that

were conducted using data collected as part of the NSP 1993 pilot test of mathematics tasks at Grades 4 and 8.

Generalizability Study

The design of the generalizability studies for the NSP pilot tasks called for the administration of multiple tasks to all students in a generalizability study (G-study) sample and the scoring of each student's responses by two or more raters. It was neither desirable nor feasible to administer all pilot mathematics tasks to a given sample of students. Instead, bundles of tasks were constructed such that all the tasks in a bundle could be completed in six or fewer 1-hour class periods. Each bundle was made up of two to four tasks. Bundles were assigned to subsamples of students with the intention of having students complete all the tasks within a given bundle. Responses to tasks for a given randomly selected bundle were scored by two or more raters. Missing data and difficulties in matching student responses to different tasks resulted in the loss of some bundles for purposes of the G-study analyses and small sample sizes for the remaining bundles that were used in the analyses reported below.

The tasks included in the bundles used in the primary G-study analyses, each of which consisted of a fully crossed pupil-by-task-by-rater design, are listed in Table 1. The time required to administer each task is also shown in Table 1. The bundle and task numbers are not sequential because data matched across tasks were not available for all bundles. The originally assigned task and bundle numbers have been maintained here to facilitate cross reference to other reports or any future analyses of the pilot study data.

Each bundle of tasks or a subset of tasks in a bundle (e.g., long tasks only in bundle 7 at Grade 4 or short form tasks only in bundle 9 at Grade 4) were analyzed separately. Estimates of variance components were computed and used to project generalizability coefficients and standard errors of measurement for various combinations of number of raters and number of tasks that might be used in a decision study (see, for example, Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991 for a description of analysis procedures for a two-facet, crossed design). Since the focus in the NSP is on absolute decisions in reference to a standard of performance, emphasis is placed on estimates of standard errors of measurement for absolute decisions.

Table 1

Task Names, Administration Times, and Bundle Assignments for Generalizability Study Assignments

Grade	Bundle	Task number and name	Administration time (Hours)
4	4	(1) Mystery Bag Game	2
		(2) Fair Share	1
		(4) Design a Flag	2 to 3
4	6	(6) Corral	2 to 3
		(12) Short Form A	1
4	7	(4) Design a Flag	2 to 3
		(7) Sharing	1
		(14) Short Form C	1
		(15) Short Form D	1
4	8	(2) Fair Share	1
		(7) Sharing	1
		(9) Checkers	2
		(10) Partitions	2 to 3
4	9	(1) Mystery Bag Game	2
		(11) Entry Codes	2
		(12) Short Form A	1
		(13) Short Form B	1
4	10	(6) Corral	2 to 3
		(9) Checkers	2
8	5	(20) The Big Prize	2
		(22) Carnival Stall	2 to 3
		(31) Short Form A	1
8	8	(23) Design a Box	1
		(29) Chocolate Boxes	2
		(31) Short Form A	1
8	9	(25) Shoelaces	2
		(26) Ramping Up	2
		(30) Sports Bag	2

Due to missing data for a given task or rater, the sample size available for a G-study analysis for a given bundle of tasks varied as a function of the number of tasks and the number of raters included in the analysis. In all cases the number of pupils available for a given analysis was relatively small (ranging from a low of 30 pupils for bundle 5 at Grade 8 to a high of 63 for long tasks only in bundle 10 at Grade 4). Hence, the results obtained for a single bundle should be viewed with caution. Consistent results observed across bundles, however,

provide a good indication of the likely levels of generalizability that can be expected for assessment designs with varying numbers of raters and tasks of the type used in the pilot study.

Because of the relatively small number of students per bundle for the full pupil-by-task-by-rater design, we did some supplementary analyses for pairs of tasks ignoring rater. The supplementary analyses involved a simple pupil-by-task design and allowed the inclusion of four more pairs of tasks at Grade 4 with sample sizes ranging from 100 to 197 per pair of tasks and four more pairs of tasks at Grade 8 with sample sizes ranging from 112 to 334 per pair of tasks.

The sample of students with complete data for the G-study analyses reflected the diversity that was sought in the overall spring pilot study. At both Grades 4 and 8, approximately 3 out of 5 students (60% at Grade 4 and 58% at Grade 8) with complete data for one of the bundles used in the pupil-by-task-by-rater G-study analyses are White; 1 in 5 is African American (20% at Grade 4 and 23% at Grade 8); and the remainder are classified as Asian American (3% at each grade), Hispanic (10% at Grade 4 and 12% at Grade 8), Native American (2% at Grade 4 and 1% at Grade 8), or "other" (3% at Grade 4 and 2% at Grade 8). Data on racial/ethnic status were missing for 1% of the students at each grade. At Grade 4, 31% of the students were reported to be participating in free or reduced lunch programs, 18% in Chapter 1, and 5% in either bilingual education or an ESL program. Five percent of the Grade 4 sample were reported to have a learning disability, and 9% were reported to be participating in a gifted or talented program. The corresponding percentages at Grade 8 are: 10% free or reduced lunch, 6% Chapter 1, 12% bilingual education or ESL, 4% with a learning disability, and 16% gifted or talented program.

Results

The results are first presented separately for the long tasks and for the short forms. Results are then presented for analyses of bundles that included a combination of long tasks and short forms.

Long tasks. Analyses of long tasks only were conducted for a total of 6 bundles (5 at Grade 4 and 1 at Grade 8). The bundles, the number of pupils with complete data available for each analysis, and the estimated variance components from the analyses of long tasks are listed in Table 2. To facilitate

Table 2

Estimated Variance Components for Analyses of Long Tasks

Grade:Bundle	G4:B4	G4:B7	G4:B8	G4:B9	G4:B10	G8:B9
Source of variance	Variance component[a]					
Pupils	.110	.107	.343	.212	.194	.254
Tasks	.118	.119	.270	.207	.245	.393
Raters	.000	.000	.000	.015	.007	.001
Pupil by Task	.382	.472	.356	.377	.215	.235
Pupil by Rater	.000	.056	.000	.023	.000	.008
Task by Rater	.015	.039	.018	.000	.032	.013
PTR, error	.297	.232	.161	.203	.245	.113
Number of pupils	40	61	49	54	63	37

[a] Variance components with negative estimates have been set to .000.

comparisons across components of variance and consistencies within each component across bundles, the components listed in Table 2 are also shown graphically in Figure 1. As can be seen in Figure 1, the variance components associated with raters and with the pupil-by-rater and task-by-rater interactions were all small compared to the variance components associated with the remaining sources of variance. Thus, there appears to be little measurement error attributable to differences in rater stringency (the rater component) or to differential responses of raters to particular pupils or tasks (the person-by-rater and task-by-rater components). Instead, measurement error due to lack of perfect rater consistency is mostly contained in the confounded three-way interaction and error component of variance.

Consistent with results of generalizability studies of other performance-based assessment tasks (e.g., Dunbar, Koretz, & Hoover, 1991; Linn, 1993; Shavelson, Baxter, & Gao, 1993), the errors of measurement attributable to the sampling of tasks (see the task and pupil-by-task components) are consistently larger than the errors attributable to raters. These results indicate that measurement error can be reduced more rapidly and effectively by increasing the number of tasks than by increasing the number of raters.

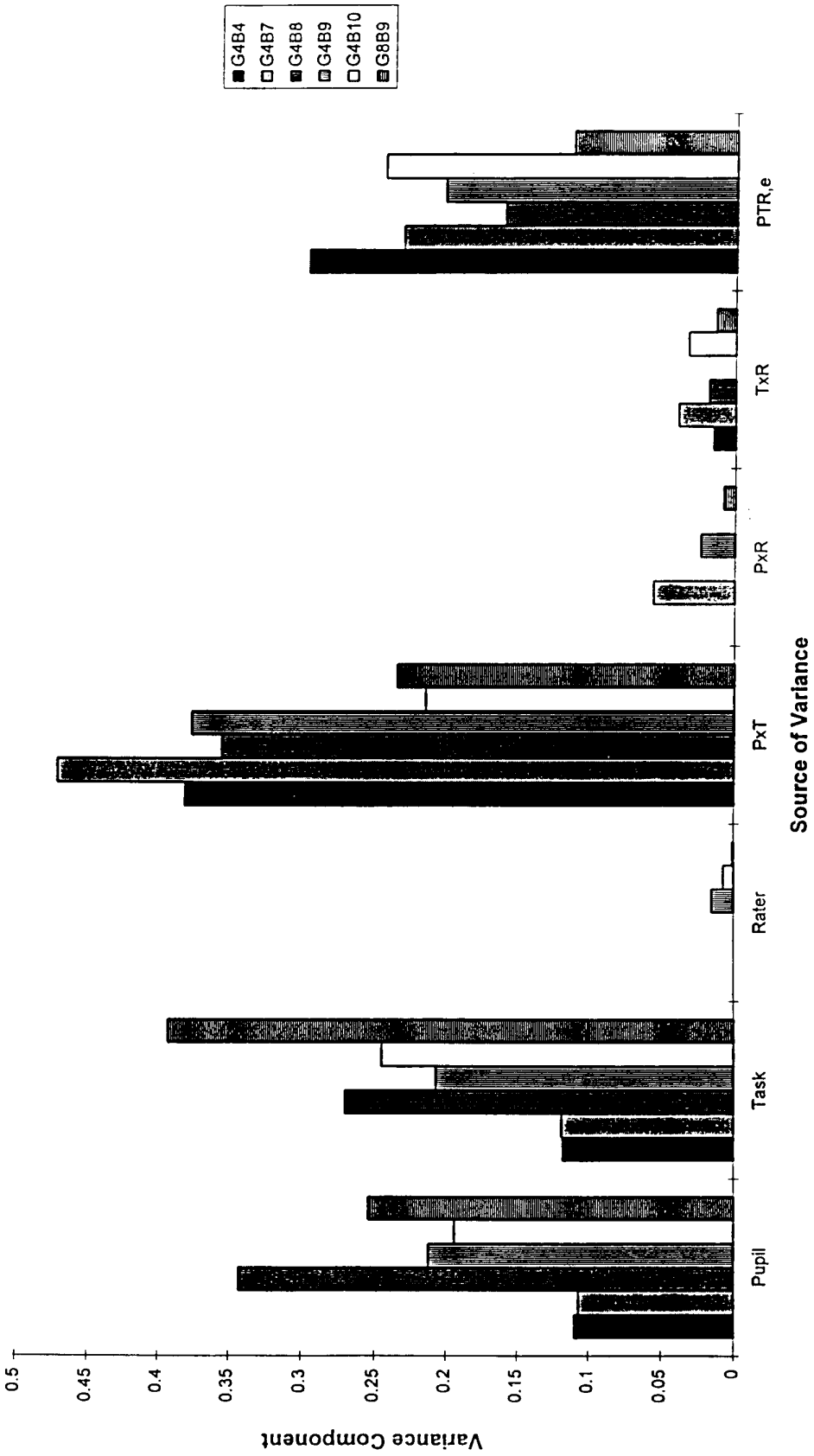

Figure 1. Components of variance for six bundles of long tasks.

The variance components contributing to measurement error in a decision study are listed in Table 3 together with the symbol used to denote each component. Also listed in the third column of Table 3 are the divisors associated with each variance component. The variance components shown in column 2 of the table are divided by the terms shown in the third column and those results are summed to produce the variance of the errors of measurement for a design with a given number of raters and a given number of tasks. The three large variance components from the results in Table 2 are associated with task, person-by-task, and the confounded person-by-task-by-rater and error. The first two of these large components of error are divided by the number of tasks while the third term is divided by the product of the number of tasks and the number of raters. Since increases in the number of tasks lead to decreases in all three of the large contributors to error while increases in number of raters contribute to a decrease in only one of the three large components, it is clear that increasing

Table 3

Contributions of Variance Components to Absolute Measurement Error Variance (σ^2_Δ)

Source of variance	Variance	Divisor for contribution to variance of measurement error
Task (T)	σ^2_T	Number of tasks (n_T)
Rater (R)	σ^2_R	Number of tasks (n_R)
Pupil by Task (PT)	σ^2_{PT}	n_T
Pupil by Rater (PR)	σ^2_{PR}	n_R
Task by Rater (TR)	σ^2_{TR}	$n_T n_R$
PTR, error	$\sigma^2_{PRT,E}$	$n_T n_R$

Variance of errors of measurement:

$$\sigma^2_\Delta = \sigma^2_T / n_T + \sigma^2_R / n_R + \sigma^2_{PT} / n_T + \sigma^2_{PR} / n_R + \sigma^2_{TR} / (n_T n_R) + \sigma^2_{PRT,E} / (n_T n_R)$$

the number of tasks will contribute more to a reduction in the magnitude of measurement error than will a corresponding increase in the number of raters.

The absolute generalizability coefficient for any given number of tasks and number of raters can be estimated by dividing the variance component for pupils by the sum of that component plus the variance due to errors shown at the bottom of Table 3. The resulting generalizability coefficients based on the data from bundle 4 at Grade 4 are displayed in Figure 2 for one and two raters as a function of the number of tasks. As can be seen, the gain in generalizability due to increasing from one to two raters is small in comparison to the gain from increasing from one to two tasks. It is also apparent from Figure 2 that a substantial number of tasks is needed to achieve a generalizability coefficient as high as .70. Even with 15 tasks each rated by two raters, the generalizability coefficient is less than .80, a value that some would consider a minimum for purposes of making important decisions about individual pupils.

The low generalizability for the tasks in bundle 4, Grade 4 is due in part to the large error components and in part to the relatively small variance component due to pupils. Four of the remaining five bundles have larger variance components associated with pupils (see Figure 1) and therefore display better generalizability despite the sampling variability due to task. This is evident in Figure 3 where the generalizability coefficients of all six bundles with variance components summarized in Table 2 are plotted as a function of the number of tasks for the situation where task responses are scored by two raters. Indeed, Grade 4, bundle 4 has the lowest generalizability of any of the six bundles shown in Figure 3. Even the bundle with the highest level of generalizability, bundle 7 at Grade 4, requires 8 tasks with two raters to achieve a generalizability coefficient of .80, however.

Although the generalizability coefficients provide useful information, they are less relevant than the standard error of measurement to an overall evaluation of the dependability of the assessment. The standard error of measurement is simply the square root of the variance of the error of measurement shown at the bottom of Table 3. The standard error of measurement is particularly relevant in the NSP situation because it provides the basis for constructing confidence intervals. Assuming that errors of measurement (not necessarily observed or universe scores) are normally distributed, one could be 95% confident that an observed score for a given

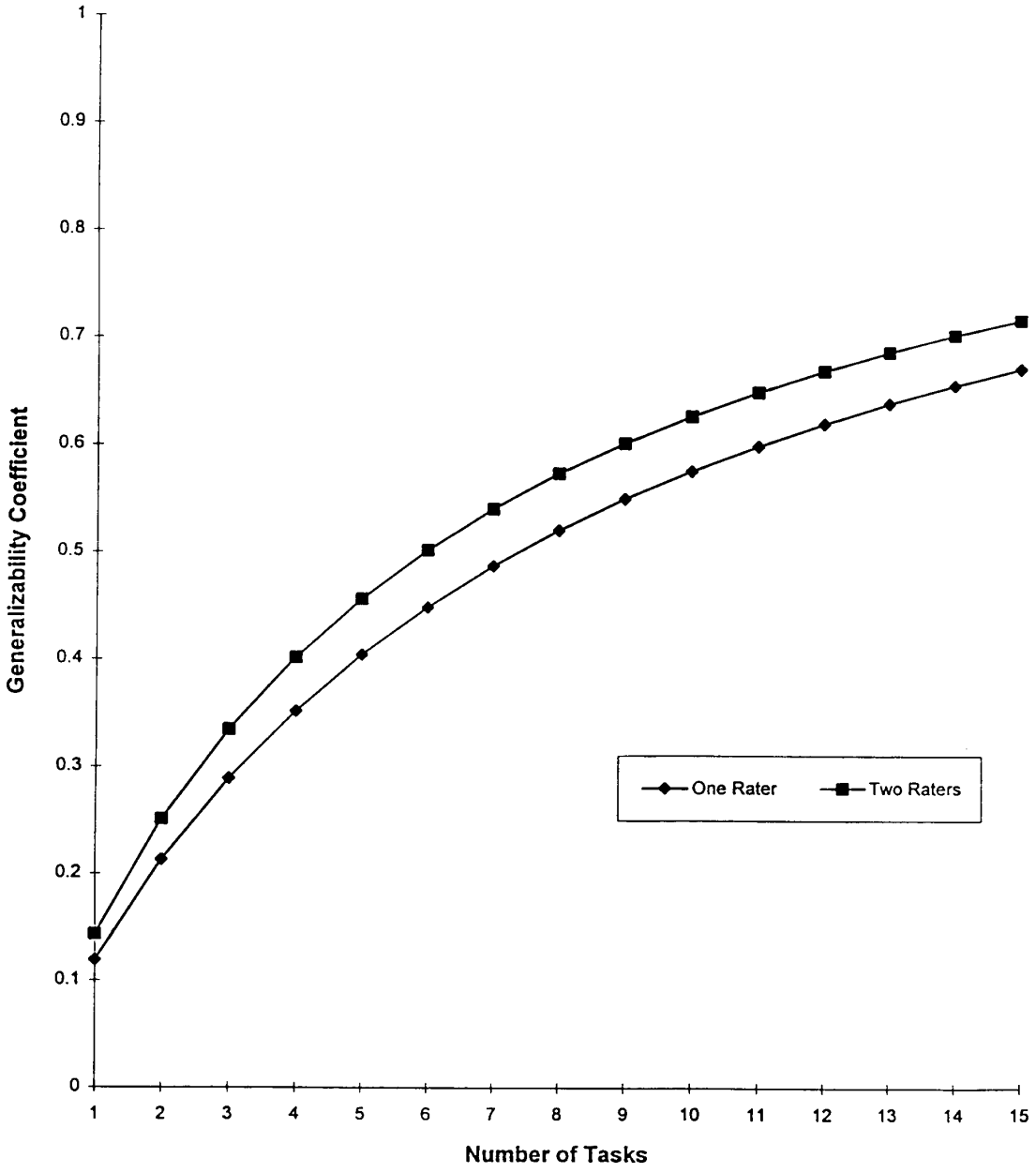

Figure 2. Generalizability coefficients for absolute decisions as a function of number of raters and number of tasks (based on Grade 4 bundle 4 data).

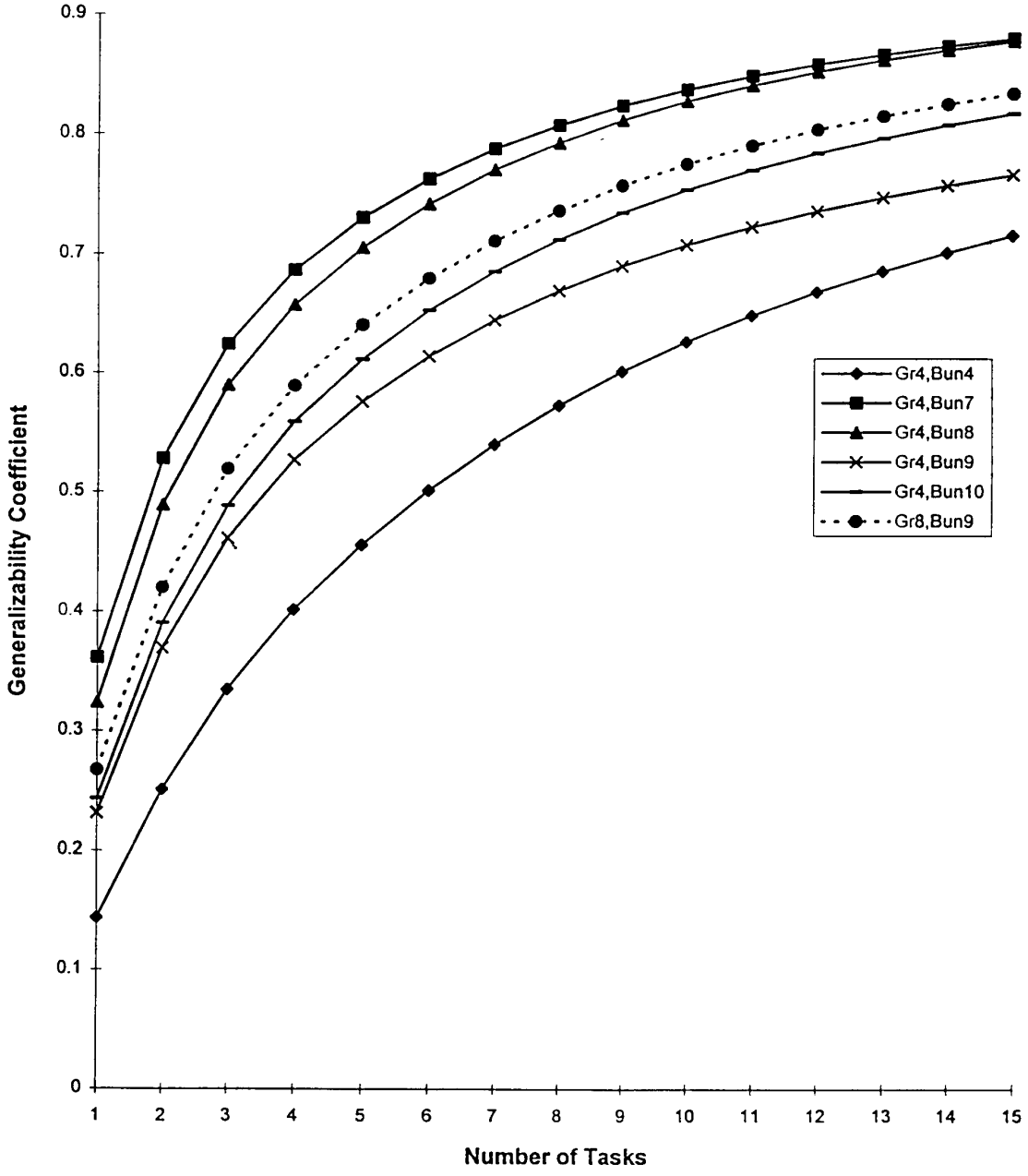

Figure 3. Generalizability coefficients for absolute decisions with two raters as a function of the number of tasks (based on six bundles of long tasks).

number of tasks and raters would be within a range of plus or minus two standard errors of measurement of a pupil's universe score.

By adding and subtracting two standard errors of measurement to a passing standard, an upper bound can be defined, above which fewer than 2.5% of pupils with universe scores equal to the standard would be expected to score, and a lower bound can be similarly defined, below which fewer than 2.5% of such students would be expected to score. Thus, pupils with scores two standard errors or more below the NSP standard can confidently be said to have failed to meet the NSP standard. Similarly, pupils with scores two standard errors or more above the standard can be confidently said to have met the NSP standard. Pupils with scores less than two standard errors of measurement above or below the NSP standard are in an uncertain range where the pass/fail decision cannot be made with 95% confidence.

The standard error of measurement decreases with increases in the number of raters or increases in the number or tasks. As was true of generalizability coefficients, however, the relative effect on the standard error of increasing the number of tasks is greater than that of increasing the number of raters. Standard errors of measurement for one and two raters are displayed in Figure 4 as a function of the number of tasks for the Grade 4, bundle 4 data. As expected the effect of changes in the number of tasks is greater than the effect of changes in the number of raters. To achieve a 95% confidence interval of 1 full score point or less on the 4-point scale used to score these tasks the standard error of measurement would have to be less than or equal to .25. For the case of the Grade 4, bundle 4 results a minimum of 11 tasks each scored by 2 raters or 13 tasks each scored by a single rater would be required to obtain a standard error of measurement of less than .25.

Figure 5 displays the standard error of measurement as a function of the number of tasks for the case of two raters per task for all six bundles for which variance component estimates for long tasks were presented in Table 2. As can be seen, with the exception of Grade 4, bundle 9, the plots of the standard errors of measurement are quite similar. The greater similarity of the standard error of measurement results shown in Figure 5 in comparison to the generalizability results shown in Figure 3 for these same data is due, in part, to the influence of the difference in the magnitude of the variance component for pupils, which

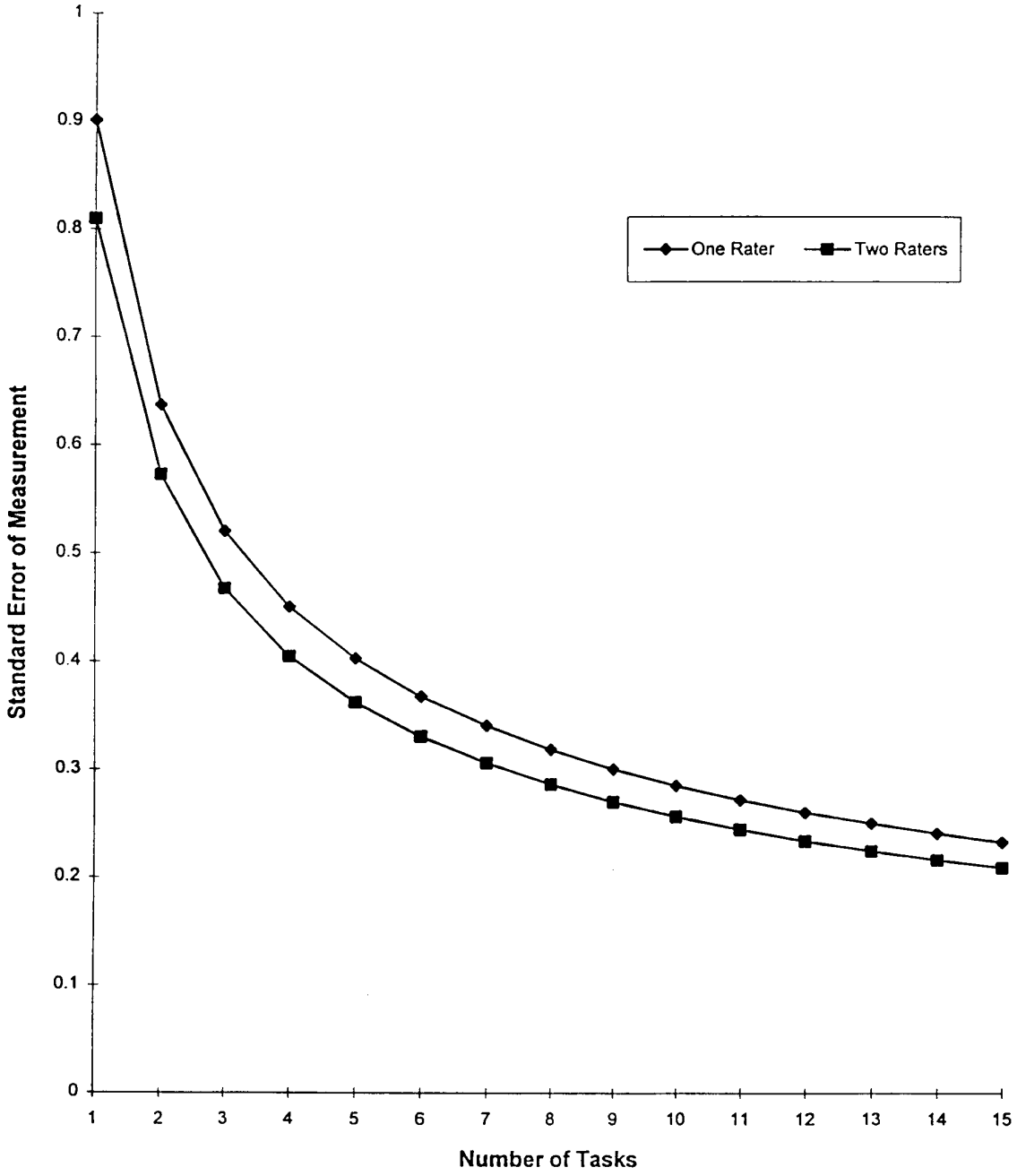

Figure 4. Standard error of measurement for absolute decisions as a function of number of raters and number of tasks (based on Grade 4 bundle 4 data).

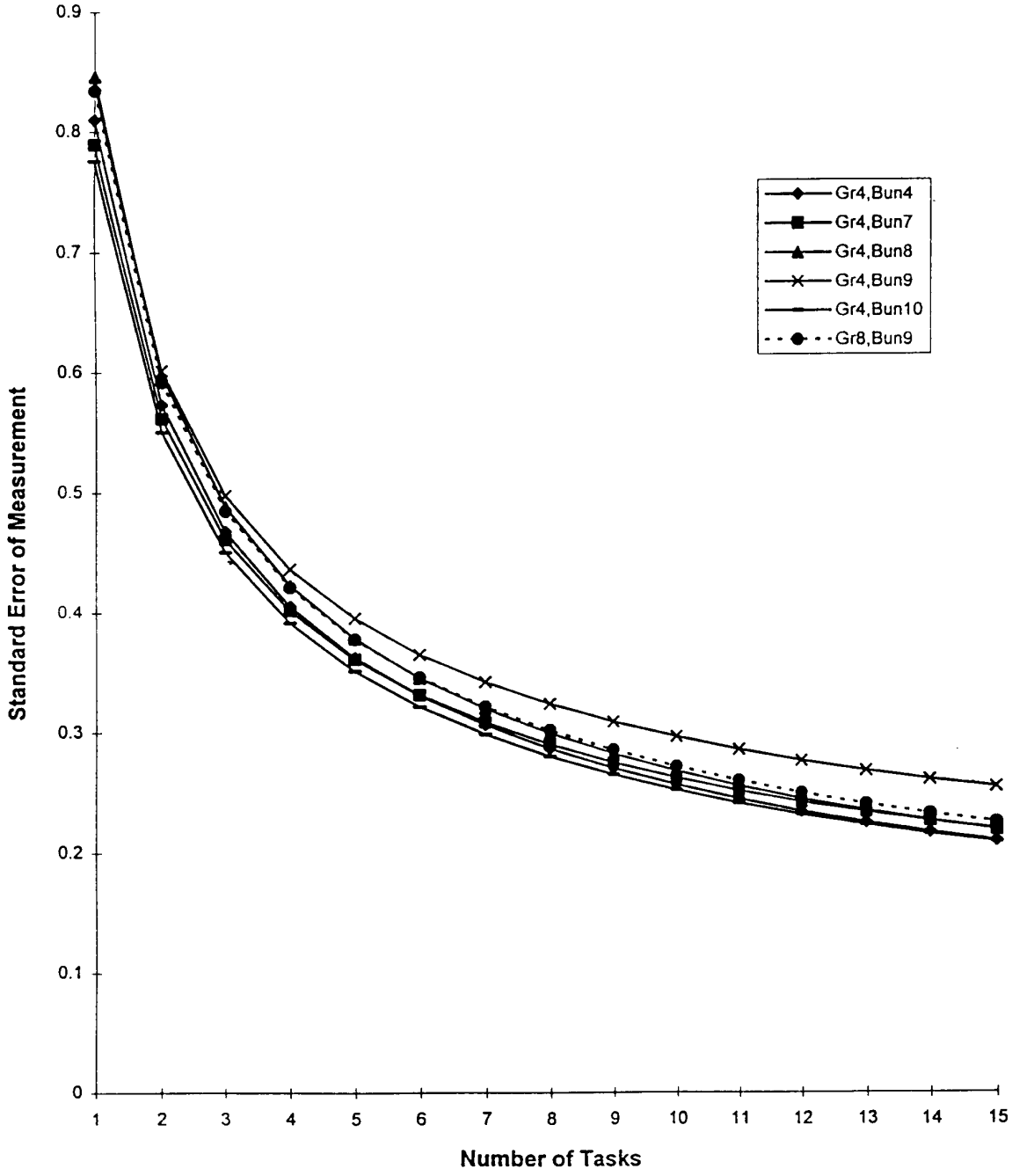

Figure 5. Standard error of measurement for absolute decisions with two raters as a function of the number of tasks (based on six bundles of long tasks).

influences generalizability coefficients but does not influence the standard error of measurement, and, in part, to the difference in metrics. The standard error of measurement is expressed in terms of the scale of measurement, whereas the generalizability coefficient is based on a ratio of variances.

With two raters per task, the number of tasks that would be required to yield a 95% confidence interval of 1.0 or less (or equivalently, a standard error of measurement of .25 or less) was either 11 or 12 tasks for all but bundle 9, Grade 4. In the case of the latter bundle, the corresponding number of tasks is 16.

Results from the supplementary analyses of six pairs of long tasks are summarized in Table 4. These results are based on simple pupil-by-task analyses and ignore raters as a source of variation. A score from a single rater was obtained for each student on each task and different raters typically scored the responses of a given student to different tasks. Listed in Table 4 are the grade, the task numbers (with the exception of task 21, see Table 1 for task names corresponding to these numbers), the number of students with scored

Table 4

Supplementary Analyses for Pupil-by-Task Designs (Ignoring Rater) Based on Combinations of Two Long Tasks

Grade	4	4	8	8	8	8
Pair of tasks	1 & 11	6 & 9	21 & 29	23 & 29	25 & 26	25 & 30
Number of students	118	100	112	285	334	169
Source of variance	Estimated variance components[a]					
Pupil	.157	.212	.298	.069	.111	.083
Task	.447	.084	.000	.075	.523	.316
PT, error	.358	.436	.554	.983	.495	.480
Estimated number of tasks required to obtain a standard error of measurement less than or equal .25	13	9	9	17	17	13

[a] Variance component with negative estimate has been set to .000.

responses on both tasks in a pair, the three components of variance (pupil, task, and the confounded pupil-by-task and error component), and the estimated number of tasks required to yield a standard error of measurement of .25 or less. Task 21 is a number sense and operations problem called "Sweet Persuasion" and has a two-hour administration time. The last two pairs of Grade 8 tasks in Table 4 (i.e., tasks 25 and 26 and tasks 25 and 30) are both part of bundle 9, and 105 pupils are involved in both analyses.

As can be seen, estimates of the number of long tasks required to obtain a standard error of measurement of .25 range from 9 to 17 tasks. These numbers are reasonably similar to the estimates based on the pupil-by-task-by-rater analysis and add strength to the general conclusions regarding the need for a sizable number of tasks to achieve dependable scores for making decisions about individual pupils.

Short forms. Analyses of the dichotomously-scored tasks contained in the short forms were conducted. Because of possible interdependencies among the short tasks contained in a given short form due to context or cueing effects, however, we report here only the analyses of the pairs of short forms contained in bundles 7 and 9 at Grade 4. The scores used in these analyses are the sum of the 4 short form tasks on each short form. The variance components for the short forms in these two bundles are listed in Table 5.

Table 5
Estimated Variance Components for Analyses of Short Forms

Grade:Bundle	G4:B7	G4:B9
Source of variance	Variance component[a]	
Pupils	.637	.842
Forms	.000	.137
Raters	.000	.008
Pupil by Form	.447	.402
Pupil by Rater	.000	.000
Form by Rater	.007	.001
PFR, error	.131	.293
Number of pupils	58	51

[a] Variance components with negative estimates have been set to .000.

The most notable difference between the variance components for the short forms shown in Table 5 and those for long tasks shown in Table 2 is in the size of the variance component for pupils. The variance component for pupils is considerably larger for the short forms than for the long tasks. The difference in the size of the variance component for pupils for the two types of tasks has a major effect on the magnitude of the generalizability coefficients, but does not influence the magnitude of the standard errors of measurement.

As was true of long tasks, the variance components for raters and for the interactions of pupil with rater and form with rater are quite small for the short forms. The pupil-by-form interaction is the largest contributor to measurement error for the short forms in both bundles.

Generalizability coefficients for the short forms based on two raters per form are plotted in Figure 6 as a function of the number of short forms. The corresponding plots of the standard errors of measurement are shown in Figure 7. As can be seen, the plots of the generalizability coefficients are almost identical. For either bundle it is estimated that an absolute generalizability coefficient exceeding .80 would require a minimum of 4 short forms (see Figure 6). To achieve a 95% confidence interval less than 1.0 (standard error of measurement less than .25) a minimum number of short forms would be 9 based on the bundle 7 results and 12 based on the bundle 9 results (see Figure 7).

Bundles with a combination of long tasks and short forms. There were five bundles that contained both long tasks and short forms. Matrices of intercorrelations of ratings provided by individual raters to each task within a bundle were computed and comparisons were made between the correlations among long tasks, those between short forms, and of long tasks with short forms. The correlation matrices, mean ratings, and standard deviations for the two bundles that contained two long tasks and two short forms are reported in Tables 6 and 7.

The small triangles of correlations on the major diagonal of Tables 6 and 7 report the interrater correlations for a single task or short form. The three-by-three boxes off the main diagonal in the tables show either the intercorrelations of one long task with another for the three raters, the intercorrelations of one short form with another, or the intercorrelations of a short form with a long task.

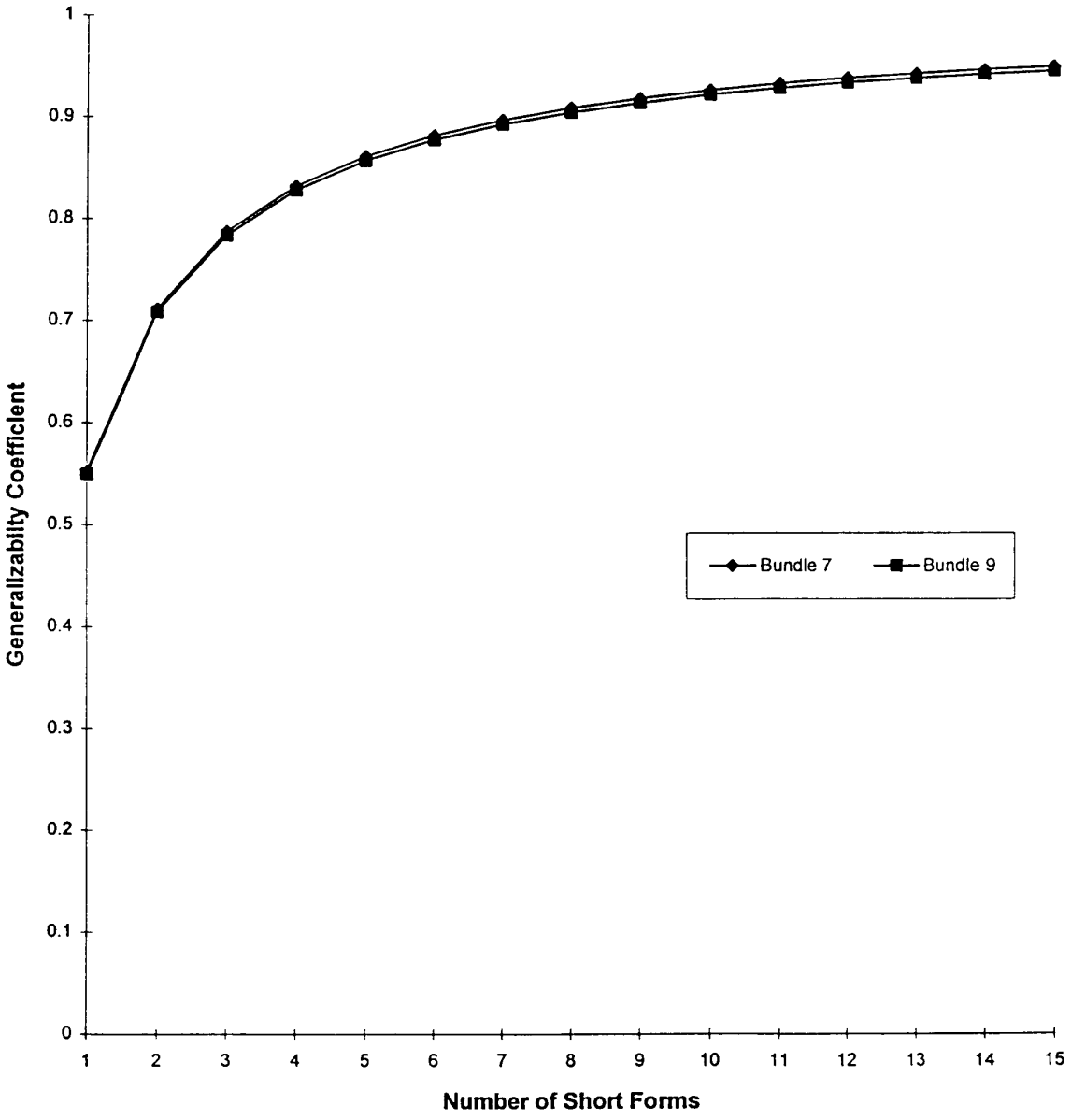

Figure 6. Generalizability coefficients for absolute decisions with two raters as a function of the number of short forms (based on two bundles of short forms).

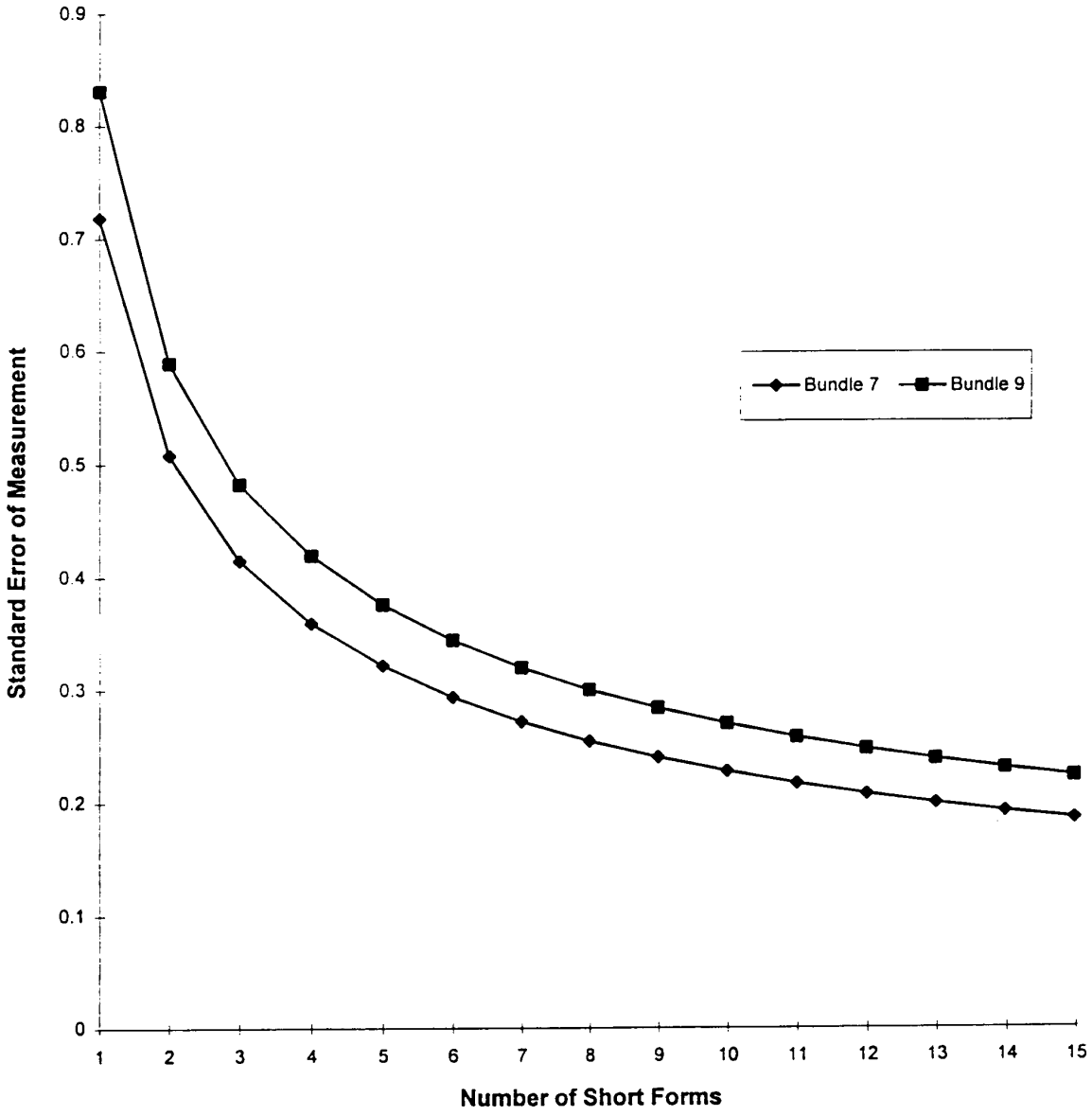

Figure 7. Standard error of measurement for absolute decisions with two raters as a function of the number of short forms (based on two bundles of short forms).

Table 6

Means and Intercorrelations of Ratings From Three Raters on Two Long Tasks and Two Short Forms in Bundle 7, Grade 4 (N = 54)

Task	Rater	4			7			14			15		
		1	2	3	1	2	3	1	2	3	1	2	3
4	1	1.0											
	2	.44	1.0										
	3	.45	.57	1.0									
7	1	.09	.16	.01	1.0								
	2	.16	.26	-.05	.79	1.0							
	3	.06	.20	.18	.81	.72	1.0						
14	1	.21	.50	.35	.25	.40	.31	1.0					
	2	.26	.45	.35	.18	.33	.24	.95	1.0				
	3	.26	.47	.39	.17	.27	.25	.91	.92	1.0			
15	1	.24	.48	.45	.30	.39	.38	.57	.56	.53	1.0		
	2	.25	.49	.44	.43	.49	.45	.64	.59	.52	.91	1.0	
	3	.16	.44	.47	.35	.41	.41	.58	.54	.48	.84	.88	1.0
Mean		1.9	2.2	2.0	2.7	2.5	2.8	2.5	2.4	2.5	2.4	2.5	2.7
Standard deviation		.7	.9	.8	1.1	1.0	1.1	.9	.9	.9	1.3	1.2	1.3

Note. Tasks 4 and 7 are "long tasks" and 14 and 15 are short forms scored as the sum of the four dichotomous short tasks.

The minor diagonals of the three-by-three boxes in Tables 6 and 7 present the correlations for a single rater when rating different tasks or short forms.

The interrater correlations are generally higher for the two short forms than they are for the two long tasks in both bundles 7 and 9. This result is in keeping with the interrater agreement findings reported by DeStefano (1993a). Correlations of the ratings of one short form with another are consistently higher than the correlations of ratings of one long form with another. The correlations of ratings of a short form with ratings of a long task generally fall in between the values for one long task with another and those for one short form with another. Those results suggest that performance on a short form tends to be a better predictor of performance on a long task than the performance on one long task is of the performance on another long task. In any event, the correlations do not

Table 7

Means and Intercorrelations of Ratings From Three Raters on Two Long Tasks and Two Short Forms in Bundle 9, Grade 4 (N = 44)

Task	Rater	1			11			12			13		
		1	2	3	1	2	3	1	2	3	1	2	3
1	1	1.0											
	2	.48	1.0										
	3	.80	.50	1.0									
11	1	.27	.06	.32	1.0								
	2	.25	.16	.31	.76	1.0							
	3	.22	.08	.26	.83	.82	1.0						
12	1	.49	.29	.47	.45	.46	.50	1.0					
	2	.49	.20	.42	.49	.52	.50	.82	1.0				
	3	.29	.15	.30	.35	.34	.45	.69	.72	1.0			
13	1	.43	.25	.39	.46	.49	.43	.45	.61	.50	1.0		
	2	.50	.34	.44	.40	.41	.36	.51	.62	.53	.90	1.0	
	3	.45	.20	.37	.43	.40	.42	.43	.60	.50	.96	.88	1.0
Mean		1.9	1.8	1.7	2.5	2.4	2.2	3.5	3.6	3.3	3.0	3.0	2.9
Standard deviation		.9	.7	.7	1.0	1.0	1.0	1.3	1.4	1.2	1.2	1.1	1.2

Note. Tasks 1 and 11 are "long tasks" and 12 and 13 are short forms scored as the sum of the four dichotomous short tasks.

suggest that the short forms are measuring something distinctly different than is being measured by the long tasks. Hence, an investigation of the generalizability of bundles containing both long tasks and short forms is reasonable.

The estimated variance components for all five bundles containing both long tasks and short forms are reported in Table 8. Two of these (Grade 4, bundles 7 and 9) were included in the analyses of long tasks only and of short forms only. The remaining three bundles with both long tasks and short forms (Grade 4, bundle 6 and Grade 8, bundles 5 and 8) were not included in the analyses reported above.

Table 8

Estimated Variance Components for Analyses of Bundles With a Combination of Long Tasks and Short Forms

Grade:Bundle	G4:B6	G4:B7	G4:B9	G8:B5	G8:B8
Source of variance	Variance component[a]				
Pupils	.377	.353	.470	.220	.246
Tasks	.702	.050	.548	.309	.000
Raters	.000	.000	.008	.000	.000
Pupil by Task	.398	.461	.474	.226	.530
Pupil by Rater	.000	.014	.008	.005	.000
Task by Rater	.005	.016	.000	.007	.022
PTR, error	.145	.195	.246	.139	.197
Number of pupils	30	54	44	30	34

[a] Variance components with negative estimates have been set to .000.

Possibly the most notable difference between the variance components for the combination of long tasks and short forms from those reported in Table 2 for long tasks only is the large variance component for tasks/forms shown for two of the bundles (Grade 4, bundles 6 and 9) in Table 8. Grade 4, Short Form A (denoted as Grade 4, task 12) was included in both bundles 6 and 9. That form had a substantially higher mean score than any of the long tasks or any of the other short forms (see, for example, the mean ratings reported in Tables 6 and 7). The inclusion of the easier Short Form A led to the unusually high variance component due to tasks/forms for bundles 6 and 9 at Grade 4.

The generalizability coefficients and standard errors of measurement for the case of two raters are displayed for the five bundles with both long tasks and short forms in Figures 8 and 9, respectively. Based on the estimates shown in Figure 8, a combined total of between 8 and 13 long tasks and short forms would be required to achieve an absolute generalizability coefficient greater than or equal to .80. The estimated combined total number of long tasks and short forms needed to achieve a standard error of measurement less than or equal to .25 is 11 or 12 for Grade 4, bundle 7 or for either of the Grade 8 bundles. For the two bundles containing Short Form A, however, the combined total number of long

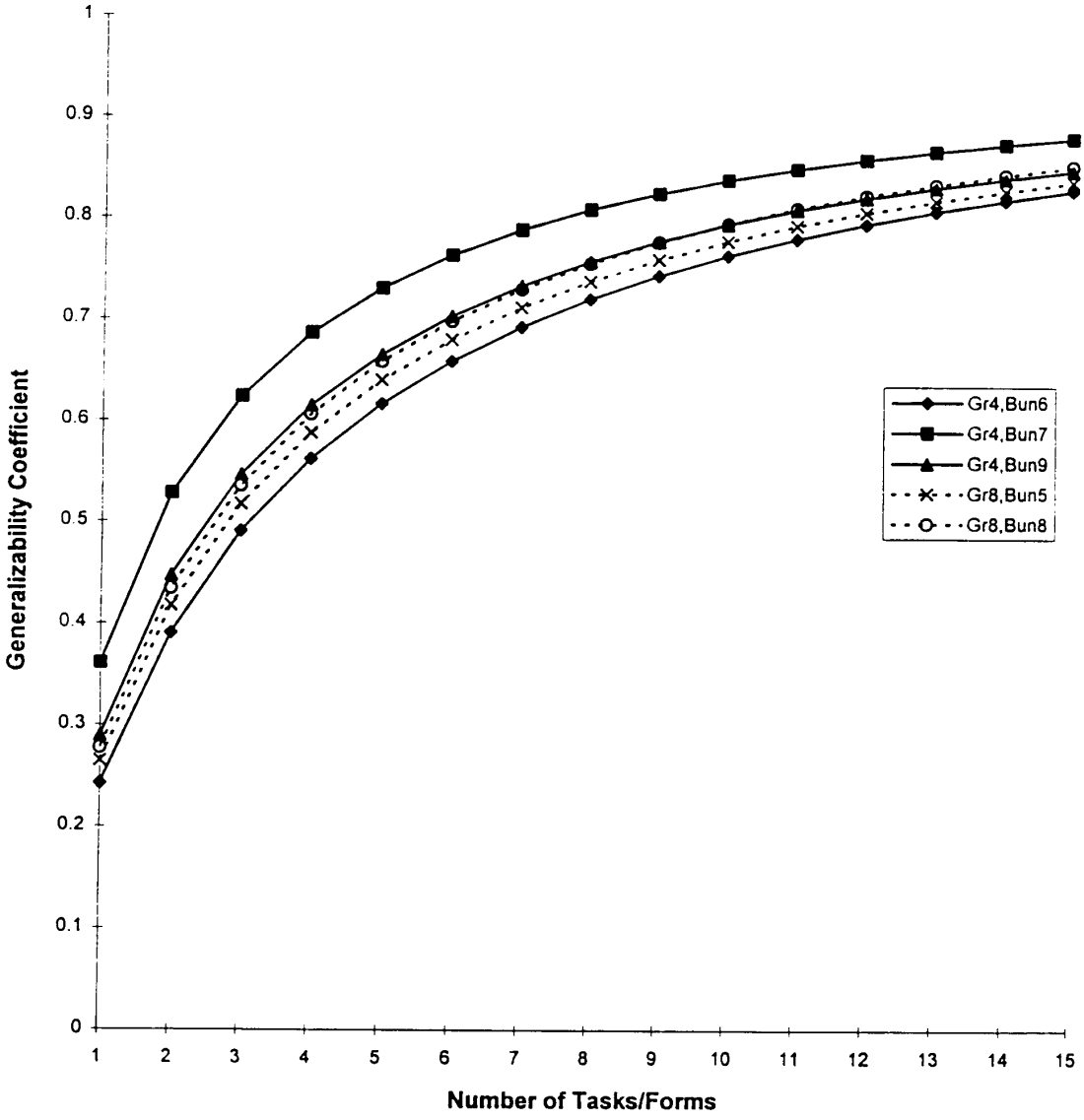

Figure 8. Generalizability coefficients for absolute decisions with two raters as a function of number of tasks/forms (based on five bundles with a combination long tasks and short forms).

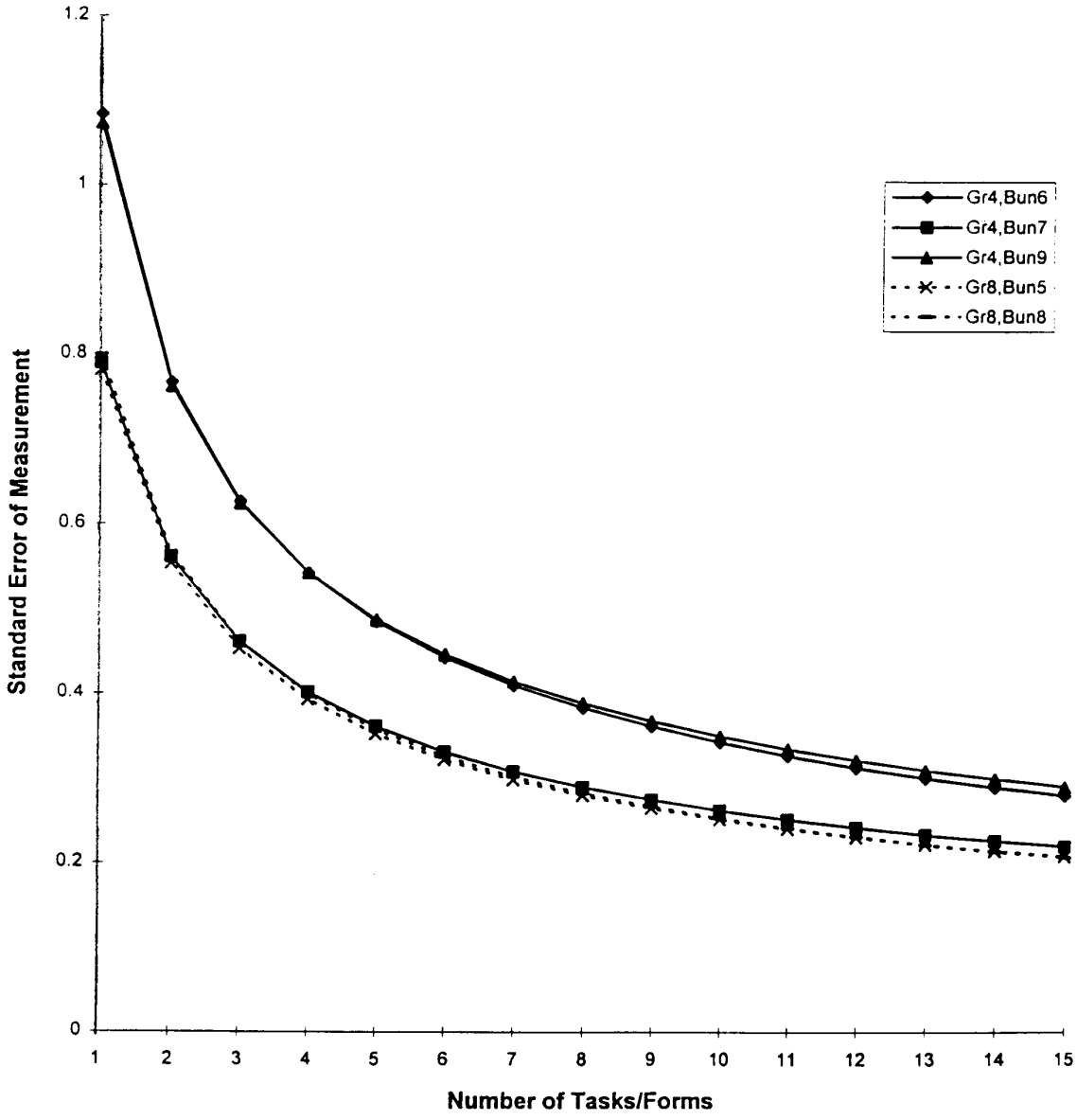

Figure 9. Standard error of measurement for absolute decisions with two raters as a function of the number of tasks/forms (based on five bundles with a combination of long task and short forms).

tasks and short forms needed for a standard error of measurement less than or equal .25 is estimated to be either 19 or 22.

Results from the supplementary analyses where rater is ignored are summarized for the pairings of long task 11 with short tasks 12 and 13 in Table 9. Twenty-five of the 103 pupils with data for the task 11 and task 12 pair also had data for task 13 and are included among the 197 pupils for the task 11 and task 13 pair. The larger estimated number of tasks (25) needed to obtain a standard error of measurement of .25 or less in Table 9 is for a pair that includes Short Form A. The smaller estimated number of tasks (15) is for a pair that includes Short Form B. Hence, these results provide general support for the primary analyses involving smaller sample sizes for the complete pupil-by-task-by-rater design.

Discussion

Although the sample size available for any given bundle is small, the G-study results obtained for the nine NSP mathematics bundles in this study are quite consistent with each other. The results are also in keeping with other investigations of generalizability for performance-based tasks (e.g., Dunbar et

Table 9

Supplementary Analyses for Pupil-by-Task Designs (Ignoring Rater) Based on Combinations of Long and Short Tasks

Grade	4	4
Pair of tasks	Long Task 11 and Short Form A	Long Task 11 and Short Form B
Number of students	103	197
Source of variance	Estimated variance components	
Pupil	.268	.544
Task	.801	.176
PT, error	.757	.718
Estimated number of tasks required to obtain a standard error of measurement less than or equal .25	25	15

al., 1991; Linn, 1993; Shavelson et al., 1993). Thus, there is strong support for the conclusions that there is considerable sampling variability due to tasks and tasks contribute much more to errors of measurement than do raters in the NSP mathematics assessment.

As was previously stated, the standard error of measurement is more relevant than a generalizability coefficient to the type of assessment use planned for the NSP. Recent work by Rogosa (1994), however, provides simulations of misclassification errors as a function of reliability coefficients that are relevant to the interpretation of generalizability coefficients obtained for the NSP tasks. He demonstrates that the misclassification error rate is quite high with a reliability of .80. These results suggest that there is a need to estimate likely levels of classification errors (false positives, i.e., the certification of pupils whose universe scores do not meet the NSP standard, and false negatives, i.e., the failure of pupils whose universe scores meet or exceed the NSP standard).

The implication of the large contributions to the standard error of measurement of task, pupil-by-task, and the confounded three-way interaction and error components implies that a sizable number of tasks is needed to make dependable decisions about individual students. Since each task or short form requires an hour or more to administer, a strategy needs to be developed either for combining some shorter tasks with long tasks or for collecting information about student performances over more extended periods of time. Both strategies are being pursued in the NSP. The plan for the 1994 mathematics reference exam pilot includes a combination of long tasks each requiring an hour to administer, medium tasks requiring 15 minutes each, and short tasks requiring 5 minutes each for a total of two hours of assessment time. Plans also call for the collection of student performance data through portfolios of work accumulated throughout the school year.

Improvements in the generalizability results for short forms might also be achieved by using more refined scores than the dichotomous right-wrong scores used for each subtask of the short forms in the pilot study. Current plans call for the use of a 3- or 4-point scale in scoring responses to 15-minute mathematics tasks from the fall 1994 administration.

Other alternatives that might improve the generalizability results include the use of sequential assessment procedures and the use of asymmetrical

decision rules. If one type of error (e.g., failing a student who should pass) is considered more serious than the other, it would be possible to use a decision rule that made fewer of that type of error at the expense of more of the opposite type of error. Sequential assessment would make it possible to use a smaller number of tasks for pupils who were far above or far below the standard while using a larger number of tasks for those who were closer to the standard.

Another intended use of NSP assessment is to provide information, not about individual pupils, but about schools. The present analyses are not directly relevant to that second use. Discussions of generalizability issues and examples where the school is the object of measurement are provided by Brennan (1993), Linn and Burton (1994), and Shavelson et al. (1993). As Brennan (1993) has shown, the generalizability for school-level aggregate results is not necessarily greater than the generalizability of individual pupil scores. The magnitude depends on the size of the variance component due to schools as well as sampling variability due to pupils within schools and school-by-task interactions. It is expected that the fall 1994 field test data will provide information on school-level generalizability for NSP tasks.

References

Brennan, R. L. (1993). *Some measurement characteristics of aggregated versus individual scores* (ACT Report Series, 93-10). Iowa City, IA: American College Testing Program.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, H. (1972). *The dependability of behavioral measurements.* New York: Wiley.

DeStefano, L. (1993a). *Summary of findings from the 1993 New Standards Project spring pilot in mathematics: Preliminary findings from Snowbird*. (Draft Technical Report, December, 21). Urbana: University of Illinois.

DeStefano, L. (1993b). *Summary of scorer qualification rates, student demographics, scorer consistency and student performance for the 1993 New Standards spring pilot.* (Draft Technical Report, December 21). Urbana: University of Illinois.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, 298-303.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, *15*, 1-16.

Linn, R. L., & Burton, E. (1994). Performance-based assessments: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*(1), 5-8, 15.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. G. Gifford & M. C. O'Conner (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Rogosa, D. (1994). *Misclassification in student performance levels* (Technical report prepared for the California Learning Assessment System). Stanford, CA: Stanford University.

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

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.