The Dependability and Interchangeability of Assessment Methods in Science

CSE Technical Report 515

Noreen M. Webb and Jonah Schlackman
CRESST / University of California, Los Angeles

Brenda Sugrue, University of Iowa

January 2000

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1 Models-Based Assessment Design: Individual and Group Problem Solving—Collaborative Assessments Noreen Webb, Project Director CRESST/UCLA

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B60002, 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 positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

THE DEPENDABILITY AND INTERCHANGEABILITY OF ASSESSMENT METHODS IN SCIENCE[1]

Noreen M. Webb and Jonah Schlackman
CRESST/University of California, Los Angeles

Brenda Sugrue, University of Iowa

Abstract

In this study, we investigated the importance of occasion as a hidden source of error variance in (a) estimates of the dependability (generalizability) of science assessment scores and (b) the interchangeability of science test formats. Two science tests were developed to measure eighth-grade students' knowledge of concepts related to electricity and electric circuits: a hands-on assessment, which provided students with equipment to manipulate, and an analogous paper-and-pencil version. Students were administered both tests on two occasions, approximately one month apart. Results of the univariate generalizability results showed that explicitly recognizing occasion as a facet of error variance altered the interpretation about the substantial sources of error in the measurement and gave lower estimates of the dependability of science scores. Including occasion as an explicit source of variance in the multivariate generalizability analyses influenced the interpretation of the observed correlation between hands-on and paper-and-pencil scores but had little influence on the estimated disattenuated correlation between assessment methods.

Performance-based assessments and other "authentic" methods of assessment have become increasingly popular as alternatives to conventional multiple-choice testing, especially in mathematics and science (National Council of Teachers of Mathematics, 1995; National Research Council, 1996). Large task-sampling variability, however, limits the reliability of scores on alternative assessments (Gao, Shavelson, & Baxter, 1994; Shavelson, Baxter, & Gao, 1993). In their study of performance-based assessments in elementary school science, Shavelson et al. (1993), for example, concluded that task-sampling variability was the major source of measurement error in performance assessments: They reported that obtaining a generalizability coefficient of .80 may necessitate administering as many as 15 tasks.

[1] We would like to thank Richard Shavelson for his helpful comments during the development of this paper. A version of this paper was presented at the annual meeting of the National Council on Measurement in Education in Montreal in April 1999.

Cronbach, Linn, Brennan, and Haertel (1997) questioned the interpretation of task-sampling variability as the major source of error in studies such as Shavelson et al. (1993) and raised the complicating issue that occasion-sampling variability may be an important *hidden* source of variance in the measurement. Because students usually perform assessment tasks only once, occasion-sampling variability is confounded with task-sampling variability. It is not known whether differences in performance across tasks on one occasion would be replicated on other occasions. If differences across tasks (both in terms of mean differences and relative ordering of examinees) remain stable across occasions, then the correct interpretation would be that task-sampling variability is a major source of error variability and occasion sampling is not. If, on the other hand, differences across tasks are not stable across occasions, then both task sampling and occasion sampling would be major sources of error variability.

Despite the possibility of large occasion-sampling variability, few empirical data are available on the stability of scores over occasions that could be used to help clarify the relative importance of task and occasion variation in assessment scores. Two studies of the stability of performance assessment scores found low to moderate stability over occasions (McBee & Barnes, 1998; Shavelson, Ruiz-Primo, & Wiley, 1999; see also Ruiz-Primo, Baxter, & Shavelson, 1993). Furthermore, their analyses disentangling the task and occasion sources of variability did, in fact, show that the task-sampling variability was due to a combination of person x task interaction and the person x task x occasion interaction (McBee & Barnes, 1998; Shavelson et al., 1999; see also Ruiz-Primo et al., 1993), showing the volatility of student performance over occasions. In the McBee and Barnes and Shavelson et al. studies of performance assessment, when data from only one occasion were analyzed, task-sampling variability was the major source of measurement error. When data from multiple occasions were analyzed, the combination of task sampling and occasion sampling proved to be the major source of measurement error.

Not only may occasion-sampling variability have an impact on conclusions about the generalizability of performance assessments; it may also have an impact on conclusions about the comparability or exchangeability of methods of testing. Given the cost of administering and scoring “authentic” performance tests, researchers have searched for less costly surrogates that yield comparable and equally valid information (see Bennett & Ward, 1993). In Shavelson et al.’s (1999)

study, considering the occasion of measurement changed their interpretation of the exchangeability of different assessment methods. Analyzing data from one occasion, they found that scores on notebook surrogates of science performance assessments were comparable to observation scores on the benchmark hands-on tasks (correlations were between .75 and .84), but that scores generated by other methods of assessment (computer simulation, multiple-choice items, and short-answer items) were not comparable (correlations ranged only from .28 to .53). When occasion was included as a source of variation, they concluded that direct observation, notebook, and computer simulation methods were equally exchangeable, but paper-and-pencil methods were not exchangeable for performance assessments.

The study reported here set out to further investigate the importance of occasion as a hidden source of error variance in science assessments. The current study used a retest design to disentangle task-sampling variability from occasion-sampling variability in performance-based and paper-and-pencil science assessments. The analyses consider the effects of occasion of measurement on estimated reliability (generalizability or dependability) and on the interchangeability of hands-on performance and paper-and-pencil scores.

Method

Study Design

Six hundred sixty-two seventh-grade and eighth-grade students (21 classes) from five Los Angeles County schools participated in the study. The schools represented a wide mix of demographic characteristics. All teachers conducted a three-week unit on electricity and electric circuits in their classrooms prior to the administration of the tests. Each teacher taught the unit using his or her usual textbook and activities; thus, instruction was not standardized across classrooms. At the end of the instructional unit, students were administered two tests: one with equipment for students to manipulate (called the hands-on test) and an analogous test in which students performed very similar tasks using pictures of the equipment instead of the equipment itself (called the pencil-and paper test; see below). The order of tests was counterbalanced within each class so that half of the students took the hands-on test on the first day while the other half took the paper-and-pencil test on the first day.

One month later, with no intervening instruction or review, students were re-administered the same two science tests (hands-on and paper-and-pencil). Students

completed the hands-on test first and the paper-and-pencil test the next day. For the hands-on test, approximately 80% of the students in each class worked on the test in collaborative three-person groups; as a control, the remaining 20% of the students worked individually at separate desks with no interaction with others. The scores on the first hands-on test were used to assign students to the collaborative group or individual condition. Consequently, the 20% of students in each class who worked individually constituted a matched sample and represented the variability of science knowledge in the class. All students completed the paper-and-pencil test individually.

The sample used in the present study consisted of the students who took the hands-on and paper-and-pencil tests individually on both occasions and who had complete data on all four tests. The final sample consisted of 57 students.[2]

Tests

Hands-on test. The hands-on test consisted of two tasks. For each task, students were given a bag of materials containing 9-volt and/or 1.5-volt batteries, wires, bulbs, and graphite resistors (Task 1: two 9-volt batteries and two 1.5-volt batteries, three bulbs in bulb holders, and seven wires with alligator clips on the ends; Task 2: two 9-volt batteries, two bulbs in bulb holders, three graphite resistors, and seven wires with alligator clips). Students were asked to assemble pairs of circuits so that the bulb in one circuit was brighter (or dimmer) than the bulb in the other circuit. After circuit construction, students were asked to draw diagrams of their circuits and answer three multiple-choice items about which of their two circuits had higher voltage, resistance, and current. Further, the test asked students to write an explanation to justify each multiple-choice answer (see Appendix A).

Students assembled a variety of circuits. Because different pairs of circuits gave rise to different correct answers on the multiple-choice and justification items, those items were scored according to the circuits that students assembled (as shown on the videotapes of their circuits). For example, if a student assembled two circuits that

[2] The 57 students analyzed here come from a larger sample of 344 students who had complete data on the hands-on and paper-and-pencil tests on the first occasion. Sugrue, Webb, and Schlackman (1998) report the results of univariate and multivariate generalizability analyses performed using the sample of 344 students (using one occasion only). Although the analyses conducted by Sugrue et al. are somewhat different in design from those used here, the results generally are quite similar to those reported here for the first occasion. The only exception is that in the larger sample of 344 students, the order of administration influenced estimated generalizability coefficients and the correlation between testing methods (for details, see Sugrue et al., 1998) while no order effects appeared in the current sample of 57 students. Consequently, the analyses here pool students taking tests in both orders.

each contained a nine-volt battery and a 1.5-volt battery, both circuits would have equal voltage. If a student assembled one circuit with one 9-volt battery and two 1.5 volt batteries and the other with one 9-volt battery, the first circuit would have a higher voltage than the second.

Each multiple-choice item received a score of correct or incorrect (0,1) depending on the circuits assembled. Justifications were scored on a 0 to 1 scale according to accuracy and completeness. For example, when asked "Why was voltage in Circuit A higher than in Circuit B?", the following scores were assigned: 1 if a student mentioned the relative number of batteries in the two circuits and the relative power or voltage generated by the batteries, 0.67 if a student mentioned the relative number of batteries in the two circuits but not the strength or voltage of the batteries, 0.33 if a student mentioned batteries but not the relative number or relative strength of them, and 0 if the explanation was irrelevant or if it displayed confusion over cause and effect (for example, "the voltage is higher because it is brighter"). Two raters scored all items independently. Rater, then, was a source of error in the generalizability analyses.

Paper-and-pencil test. The paper-and-pencil test consisted of items exactly analogous to the hands-on test except that students were given pictures of the equipment (batteries, bulbs, resistors) instead of actual manipulatives (see Appendix A). Otherwise, on this portion of the test, the students received identical instructions. Students were asked to construct two circuits (draw diagrams using the items given in order to make the bulb in one circuit brighter than the bulb in the other) and answer the same multiple-choice and justification questions.

Two raters scored all items. The task scores for both hands-on and paper-and-pencil tasks used for analysis were the mean scores across the three multiple-choice items and three justification items for each task.

Analysis

Univariate generalizability analyses were conducted to evaluate the dependability of the assessment methods and to examine the consistency of performance across tasks, raters, and occasions. Multivariate generalizability analyses were conducted to examine the universe-score correlations among testing methods disattenuated for measurement error. For both sets of analyses, two designs were used: (1) persons x tasks x raters for each occasion separately and (2)

persons x tasks x raters x occasions. The comparison of designs makes it possible to assess the importance of occasion as a source of variation in the measurements.

In the generalizability analyses, tasks, raters, and occasions were all treated as random. We considered the tasks, raters, and occasions sampled in this study as exchangeable with other tasks, raters, and occasions in the universe, and the intent was to generalize beyond the conditions sampled here. In addition, we treated the first administration of the hands-on test and the paper-and-pencil test as the same occasion even though the tests were administered on consecutive days, not the same day. Similarly, we treated the second administration of the two tests, also administered on consecutive days, as the same occasion. We considered the one-day time interval between the first administration of the hands-on and paper-and-pencil tests to be much smaller than the one-month time interval between the first and second administration of the tests. Furthermore, correlations between task scores across the two days of one occasion tended to be higher than the correlations between the two occasions. First, the average correlation between analogous tasks on the two tests administered on one occasion (e.g., Task 1 on the first hands-on test and Task 1 on the first the paper-and-pencil test) was higher than the average correlation between analogous tasks on the two tests administered on different occasions (e.g., Task 1 on the first hands-on test and Task 1 on the second paper-and-pencil test): .67 vs. .61. Second, the average correlation between non-analogous tests on the two tests administered on one occasion (e.g., Task 1 on the first hands-on test and Task 2 on the first paper-and-pencil test) was higher than the average correlation between non-analogous tasks on the two tests administered on different occasions (e.g., Task 1 on the first hands-on test and Task 2 on the second paper-and-pencil test): .62 vs. .53. These data support our contention that student performance was more consistent across the two days of one occasion than between occasions one month apart.

Results

The results are presented in two sections. The first section describes the results of the univariate generalizability analyses examining the role of occasion in estimates of dependability (reliability / generalizability). The second section presents the results of the multivariate generalizability analyses investigating the interchangeability of test formats.

Dependability of Assessment Methods

Table 1 gives the breakdown in scores by task, occasion, and rater for the hands-on and paper-and-pencil tests. Mean performance did not differ by rater or by task: Rater means were nearly identical and task means were very similar. Mean scores increased significantly from the first occasion to the next, consistent with a practice effect. A within-subjects analysis of variance with tasks, raters, and occasions as the independent factors showed significant results only for the occasion main effect: for hands-on scores, $F(1, 448) = 6.88$, $p = .009$; for paper-and-pencil scores, $F(1, 448) = 15.55$, $p = .0001$. The significance levels of the test statistics for all other effects in the model ranged from $p = .63$ to $p = .98$.

Table 2 presents correlations by task both within and between occasions for the hands-on and paper-and-pencil tests. Because rater agreement was nearly perfect (correlations between raters for each task ranged from .98 to 1.00), the correlations in Table 2 are not presented separately for each rater but use the average of the two raters' ratings. Correlations between tasks were all moderately high. The magnitude of the correlations was similar between tasks administered on the same occasion, between the same task administered on different occasions, and between different tasks administered on different occasions, with one exception. On the paper-and-pencil test, correlations between the two tasks performed on the same occasion (.74) were somewhat higher than correlations between tasks performed on different occasions (ranging from .57 to .72). Otherwise, the correlations did not seem to be

Table 1

Mean Hands-On and Paper-and-Pencil Scores by Task, Rater, and Occasion

	Occasion 1				Occasion 2			
	Rater 1		Rater 2		Rater 1		Rater 2	
Score	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Hands-on test								
Task 1	.46	.27	.47	.27	.53	.25	.53	.25
Task 2	.48	.29	.48	.29	.54	.25	.54	.25
Total[a]	.47	.25	.47	.25	.54	.23	.54	.23
Paper-and-pencil test								
Task 1	.43	.26	.44	.26	.53	.27	.53	.27
Task 2	.43	.27	.43	.27	.53	.29	.53	.29
Total[a]	.43	.24	.43	.25	.53	.26	.53	.26

[a]Mean score across Tasks 1 and 2.

Table 2

Correlations Among Tasks for the Hands-On and Paper-and-Pencil Tests

	Occasion 1 Task 1	Occasion 1 Task 2	Occasion 2 Task 1
Hands-on test			
Occasion 1, Task 1	—		
Occasion 1, Task 2	.58	—	
Occasion 2, Task 1	.64	.56	—
Occasion 2, Task 2	.50	.68	.67
Paper-and-pencil test			
Occasion 1, Task 1	—		
Occasion 1, Task 2	.74	—	
Occasion 2, Task 1	.68	.59	—
Occasion 2, Task 2	.72	.57	.74

differentially affected by varying task, varying occasion, or both. This pattern of results coincides with that of McBee and Barnes (1998), who also found that correlations for the same task across occasions were fairly similar in magnitude to the correlations between different tasks on the same occasions; although in their case the correlations tended to be low.

Tables 3 and 4 give the results of generalizability analyses for a persons x tasks x raters design for each occasion for the hands-on test and the paper-and-pencil test, respectively. On both occasions, for both tests, the two largest effects were attributable to the person effect (universe-score variance) and the interaction between persons and tasks. Across the four analyses, universe-score variance ranged from 58% to 74% of the total variance. Across the four analyses, the person x task interaction ranged from 26% to 42% of the total variance. For each test and each occasion, then, the relative standing of persons changed dramatically from one task to the other. The other effects, however, were negligible. The task main effect was zero, showing that average student performance was the same on both tasks, which is consistent with the means presented in Table 1. All effects associated with raters were near zero, showing that raters agreed in their ratings of persons and task difficulty, which is consistent with the rater means and correlations presented previously.

Tables 3 and 4 also show that on the hands-on test, but not the paper-and-pencil test, the level of estimated generalizability was higher on the second occasion than on the first. This result comes from a smaller estimated variance component for

Table 3

Estimated Variance Components for Hands-On Scores for the Persons x Tasks x Raters Design

	Occasion 1		Occasion 2	
Source of variation	Estimated variance component	Percentage of total variance	Estimated variance component	Percentage of total variance
Persons (*p*)	.0450	57.67	.0424	66.57
Tasks (*t*)	0[a]	0	0[a]	0
Raters (*r*)	.0000	0.00	0[a]	0
pt	.0323	41.47	.0211	33.17
pr	.0001	0.17	.0000	.00
tr	.0000	0.01	0[a]	0
ptr,e	.0005	0.68	.0002	.26
$\hat{\rho}^2$ [b]	.73	—	.80	—
$\hat{\Phi}$ [b]	.73	—	.80	—

[a]Negative estimated variance component set equal to zero.

[b]Estimated generalizability or phi coefficient for a test with two tasks and one rater.

Table 4

Estimated Variance Components for Paper-and-Pencil Scores for the Persons x Tasks x Rater Design

	Occasion 1		Occasion 2	
Source of variation	Estimated variance component	Percentage of total variance	Estimated variance component	Percentage of total variance
Persons (*p*)	.0515	73.71	.0582	73.64
Tasks (*t*)	0[a]	0	0[a]	0
Raters (*r*)	.0000	.00	0[a]	0
pt	.0183	26.16	.0207	26.27
pr	.0000	.01	.0000	.02
tr	0[a]	0	.0000	.00
ptr,e	.0001	.12	.0001	.07
$\hat{\rho}^2$ [b]	.85	—	.85	—
$\hat{\Phi}$ [b]	.85	—	.85	—

[a]Negative estimated variance component set equal to zero.

[b]Estimated generalizability or phi coefficient for a test with two tasks and one rater.

the person x task interaction on the second occasion. Students showed more consistency across the two tasks on the second administration of the hands-on test than on the first administration. The correlations in Table 2 also bear this out: for the hands-on test, the correlation between tasks on the second occasion (.67) is higher than the correlation between tasks on the first occasion (.58). For the paper-and-pencil test, in contrast, the correlations between tasks are the same on both occasions (.74).

When occasion was included as a facet in the generalizability analyses, the relative importance of the effects changed, as did estimated generalizability and dependability. The results presented in Table 5 still show a substantial effect for persons, ranging from 52% to 59% of the total variation. However, the effect of the person x task interaction is much reduced—12% for the hands-on test and 0% for the paper-and-pencil test—and the person x task x occasion interaction emerges as the largest effect other than universe-score variance—24% for the hands-on test and 27% for the paper-and-pencil test. The large person x task x occasion interaction effect shows that a student's score depended not only on the task but also on the occasion of testing. A student who did better on Task 1 than on Task 2 on the first occasion may have shown a very different pattern of scores on the second occasion.

On the hands-on test, the person x task interaction effect was reduced when occasion was added to the design, but it remained non-negligible (12%). This result shows that, averaging over the two occasions, there was some tendency for the tasks to rank order students differently. Averaged over occasions, some students found one task easier while other students found the other task easier. On the paper-and-pencil test, however, the person x task effect dropped to zero (Table 5), showing that the apparent difference across paper-and-pencil tasks in the design that ignores occasion as a facet (the substantial person x task effect in Table 4) was entirely due to the vagaries of how a student responded to a task on a specific occasion, and not due to how a student responded to particular tasks per se. Examination of students' scores supports the different magnitudes of the estimated person x task interaction in the analyses of hands-on and paper-and-pencil scores. On the hands-on test, about half of the sample (51%) consistently showed higher scores on one task than the other: 21% showed higher scores on Task 1 than on Task 2, and 30% showed higher scores on Task 2 than on Task 1. On the paper-and-pencil test, fewer students (28%) showed consistently higher scores on one task than on the other. The majority of the sample (72%) showed no consistent pattern across tasks or occasions, for

Table 5

Estimated Variance Components for Hands-On and Paper-and-Pencil Scores for the Persons x Tasks x Raters x Occasions Design

	Hands-on		Paper-and-pencil	
Source of variation	Estimated variance component	Percentage of total variance	Estimated variance component	Percentage of total variance
Persons (*p*)	.0375	51.55	.0483	59.26
Tasks (*t*)	0[a]	0	.0000	.02
Raters (*r*)	0[a]	0	.0000	.00
Occasions (*o*)	.0020	2.77	.0049	6.06
pt	.0091	12.49	0[a]	0
pr	.0000	.01	0[a]	0
po	.0061	8.41	.0066	8.05
tr	.0000	.01	.0000	.00
to	0[a]	0	0[a]	0
ro	0[a]	0	0[a]	0
ptr	.0000	.00	0[a]	0
pto	.0176	24.20	.0216	26.50
pro	.0001	.09	.0000	.02
tro	0[a]	0	0[a]	0
ptro,e	.0003	.47	.0001	.10
$\hat{\rho}^2$ [b]	.66		.73	
$\hat{\Phi}$ [b]	.63		.68	

[a]Negative estimated variance component set equal to zero.

[b]Estimated generalizability or phi coefficient for a test with two tasks, one rater, administered on one occasion.

example, performing better on Task 1 than on Task 2 on the first occasion and performing better on Task 2 than on Task 1 on the second occasion, or vice versa, or performing equally well on the two tasks on one occasion but obtaining quite different scores on the two tasks on the other occasion.

It is possible that having available the equipment to manipulate on the hands-on version helped make certain aspects of one task or the other more salient to at least some students and, consequently, helped make their responses more systematic. For example, being able to manipulate the graphite resistors on Task 2 (a key element in that task) may have helped focus students' attention on the role of

electrical resistance in this task. Evidence that this was the case comes from the number of students who mentioned resistance as a reason for why one circuit they built was dimmer than the other. More students mentioned resistance as a cause of dimness in the circuit for this task on the hands-on test (23% mentioned resistance on this item on both occasions) than on the analogous item on the paper-and-pencil test (7%).

The findings here of a large person x task x occasion interaction and a reduced person x task interaction are consistent with the results of both McBee and Barnes (1998) and Shavelson et al. (1999). In all three studies, student performance was volatile across both tasks and occasions. This was the case whether the tasks were highly similar, as in the present study, very similar as in the matched tasks in the McBee and Barnes study (the "Basketball Camp" and "Space Camp" tasks), or less similar as in the Shavelson et al. study (the "Paper Towels," "Electric Mysteries," and "Bugs" tasks) and in McBee and Barnes' analysis of all four of their tasks ("Basketball Camp," "Space Camp," "Olympics," and "Tug of War").

Two other effects concerning occasion were non-negligible. First, the non-negligible person x occasion interaction effect (8% for both the hands-on and paper-and-pencil tests) shows the tendency of the relative standing of students to change from the first occasion to the second, averaging over the two tasks. Some students did better on the first occasion, whereas others did better on the second occasion. This result shows that the general pattern of improvement from the first occasion to the second (Table 1) did not apply to all students. Second, the non-negligible occasion main effect for the paper-and-pencil test (6%) reflects the overall improvement in performance across occasions.

Tests are typically administered once and estimates of reliability (generalizability or dependability) are calculated using scores on that one occasion. To determine whether the inclusion of multiple occasions in the estimation process would change the estimates, we compared the relative and absolute generalizability coefficients ($\hat{\rho}^2$) and phi coefficients ($\hat{\Phi}$) generated by a design that did not include occasion as an explicit facet with estimates generated by a design that did include occasion as a facet. That is, we compared estimates of generalizability from a persons x raters x tasks design (using scores from Occasion 1 only), and a persons x raters x tasks x occasions design (using scores from two occasions). Because there was near-zero rater variation in this study, the estimated generalizability and phi coefficients are calculated for one rater. In Table 6, the generalizability coefficient

Table 6

Generalizability and Dependability Coefficients for Hands-on and Paper-and-Pencil Scores for Different Numbers of Tasks

	Number of tasks[a]						
	1	2	3	4	5	8	15
Hands-on test							
p x t x r design[b]							
$\hat{\rho}^2$	.58	.73	.80	.84	.87	.91	.95
$\hat{\Phi}$	.58	.73	.80	.84	.87	.91	.95
p x t x r x o design							
$\hat{\rho}^2$	.53	.66	.71	.74	.76	.80	.82
$\hat{\Phi}$	.52	.63	.69	.71	.73	.76	.79
Paper-and-pencil test							
p x t x r design[b]							
$\hat{\rho}^2$	.74	.85	.89	.92	.93	.96	.98
$\hat{\Phi}$	.74	.85	.89	.92	.93	.96	.98
p x t x r x o design							
$\hat{\rho}^2$	.63	.73	.78	.80	.82	.84	.86
$\hat{\Phi}$	.59	.68	.72	.74	.75	.77	.79

[a]Estimated coefficients are for one rater and one occasion.

[b]Analysis of Occasion 1 data only.

($\hat{\rho}^2$) shows the dependability (reliability) of the relative ordering of students (a norm-referenced interpretation of test scores), whereas the phi coefficient ($\hat{\Phi}$) shows the dependability (reliability) of the absolute level of a student's performance independent of others' performance (cf. criterion-referenced interpretation).

As can be seen in Table 6, the conclusions regarding estimated generalizability are different for the two designs. For the analyses in which occasion was not included as an explicit source of variation, estimated generalizability and dependability coefficients for a test with two tasks (the same as the 45-minute test used in the G study) were quite high for the paper-and-pencil version (.85 for both $\hat{\rho}^2$ and $\hat{\Phi}$) and still sizable for the hands-on version (.73 for both $\hat{\rho}^2$ and $\hat{\Phi}$). With three tasks, $\hat{\rho}^2$ and $\hat{\Phi}$ would be .80. When occasion is included in the design, the estimated generalizability and dependability coefficients for one occasion are considerably lower. For a 2-task test, the coefficients range from .63 to .73 across the hands-on and paper-and-pencil versions. To obtain estimated generalizability

coefficients of .80 would require 4 tasks for the paper-and-pencil version and 8 tasks for the hands-on version. Estimated dependability coefficients of .80 would require 19 tasks for the hands-on version and more than 20 tasks for the paper-and-pencil version.

The difference in estimated generalizability coefficients for the two designs arises from how the effects involving occasion are estimated in the two designs. With a persons x tasks x raters x occasions design, the effects involving occasions (e.g., the main effect for occasions, the interaction between persons and occasions, etc.) can all be estimated separately, as in Table 5. In a persons x tasks x raters design, the occasion facet is hidden. The effects involving occasion are still present, but they can not be estimated separately. Rather, they are confounded with the corresponding effects that do not involve occasions. For example, the effect for persons (*p*) is confounded with the interaction between persons and occasions (*po*). As Cronbach et al. (1997) point out, in a design that does not explicitly include occasion as a facet, such as the persons x tasks x raters design in Tables 3 and 4, "[e]ach of the seven components includes temporal effects. Thus, relabeling the pupil component as p,po would emphasize that the pupil may perform better . . . than usual by virtue of some morale-inducing event or may perform worse because of illness" (p. 384).

Because of the confounding between the *p* and *po* effects in the persons x tasks x raters design, the variance component for persons that is estimated for that design (Tables 3 and 4) is really a combination of the universe-score variance component and the variance component for the person x occasion interaction: $\sigma^2(p) + \sigma^2(po)$ A nonzero person x occasion effect, then, will lead to an overestimated universe-score variance and an underestimated error variance in the persons x tasks x raters design, producing an overestimate of the level of generalizability.

Interchangeability of Assessment Methods

The extent to which performance on science assessments is affected by having the opportunity to manipulate equipment was explored by examining mean performance and correlation of scores on the hands-on and pencil-and-paper tests. The results presented earlier in Table 1 indicated that total scores on the hands-on and paper-and-pencil tests were very similar. Thus, there was no effect of equipment manipulation on average performance. Furthermore, this result was consistent across both occasions.

Unlike mean performance scores, the correlations between hands-on and paper-and-pencil scores differed according to whether occasion was considered. Using only data from the first occasion, the observed correlation between hands-on and paper-and-pencil total scores was .69. Using data for the second occasion, the observed correlation between hands-on and paper-and-pencil total scores rose to .84. When scores were averaged over the two occasions, the correlation between hands-on and paper-and-pencil scores was .83.

The higher correlation on the second occasion and the high correlation for the average across the two occasions may have been related to a practice effect. On a repeat administration of these highly similar tests, students may have remembered their responses from one version of the test to the other or may have otherwise consolidated their performance on the two tests. Because the two tests were so similar in structure, students could have remembered or reconstructed the responses they had given on the other test. For example, on the paper-and-pencil test, students could remember or reconstruct the circuits they had made on the hands-on test on the previous day. This remembering or reconstruction of circuits would be more likely to occur on the second occasion than on the first occasion because the tests were less novel to the students on the second occasion and they would be more likely to recognize the similar structure of the two tests. An examination of the circuits that students constructed on the tests on the two occasions supports this hypothesis. On the first occasion, 40% of the students gave the same pair of circuits on Task 1 of the hands-on test as they gave on Task 1 of the paper-and-pencil test, and 56% of the students gave the same pair of circuits on Task 2 of the two tests. On the second occasion, the percentages were higher: 65% of the students gave the same pair of circuits on Task 1 of the hands-on and paper-and-pencil tests, and 77% of the students gave the same pair of circuits on Task 2 of the two tests. Because the multiple-choice and justification responses were related to students' circuits, students who gave the same pair of circuits on the two tests were more likely to generate similar multiple-choice and justification responses and, consequently, scores on the two tests, than were students who gave different pairs of circuits on the two tests.

To examine the role of various sources of error in the correlation between hands-on and paper-and-pencil scores and to obtain disattenuated correlations between hands-on and paper-and-pencil scores (corrected for unreliability in the two measures), we carried out multivariate generalizability analyses. To examine the

effects of explicitly including occasion as a facet of error in the design, we carried out multivariate generalizability analyses for two designs: persons x tasks x raters using the hands-on and paper-and-pencil scores collected on Occasion 1; and persons x tasks x raters x occasions using the scores for both occasions.

Because of the relatively infrequent use of multivariate generalizability, we will explain in detail here the statistical theory and procedures for this analysis (see also Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Webb, Shavelson, & Maddahian, 1983). In univariate generalizability theory, an observed score is decomposed into the universe score (analogous to the true score in classical test theory) and error scores corresponding to multiple, independent sources of error variation. From the analysis of variance, an estimate of each component of variation in the observed score is obtained. For example, consider the task scores on the hands-on test for Occasion 1. There were two tasks each rated by two raters. Thus, the design is *p x t x r* or, in words, persons crossed with tasks and raters. The total variance of the observed scores equals the following sum of variance components:

$$\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(tr) + \sigma^2(ptr,e) \qquad (1)$$

In Equation 1, $\sigma^2(p)$, the variance component for persons, is universe-score variation and the remaining variance components constitute error variation. Using the subscript *h* for hands-on and *p* for paper-and-pencil, the total variance of the hands-on scores and the total variance of the paper-and-pencil scores can be decomposed as follows:

$$\begin{aligned}\sigma^2(_hX_{ptr}) &= \sigma^2(_hp) + \sigma^2(_ht) + \sigma^2(_hr) + \sigma^2(_hpt) + \sigma^2(_hpr) + \sigma^2(_htr) \\ &\quad + \sigma^2(_hptr,e). \\ \sigma^2(_pX_{ptr}) &= \sigma^2(_pp) + \sigma^2(_pt) + \sigma^2(_pr) + \sigma^2(_ppt) + \sigma^2(_ppr) + \sigma^2(_ptr) \\ &\quad + \sigma^2(_pptr,e). \end{aligned} \qquad (2)$$

Multivariate generalizability theory (Brennan, 1992; Brennan, Gao, & Colton, 1995; Cronbach et al., 1972; Shavelson & Webb, 1981; Webb et al. 1983) decomposes covariances among scores as well as variances among scores. In the present study, the total covariance between hands-on and paper-and-pencil scores is decomposed into the following components of covariance:

$$\sigma(_hX_{ptr},{}_pX_{ptr}) = \sigma(_hp,{}_pp) + \sigma(_ht,{}_pt) + \sigma(_hr,{}_pr) + \sigma(_hpt,{}_ppt) + \sigma(_hpr,{}_ppr) + \sigma(_htr,{}_ptr) + \sigma(_hptr,e,{}_pptr,e) \quad (3)$$

The covariance component $\sigma(_hp,{}_pp)$ is the covariance between persons' universe scores for hands-on and paper-and-pencil. The remaining terms in Equation 3 are error covariance components. Because the same raters were used to evaluate student responses on the hands-on and paper-and-pencil tests and the tasks used on the hands-on and paper-and-pencil tests were identical with the exception of the presence of physical equipment or pictures of the equipment, the error covariance components are not equal to zero. The disattenuated correlation (see Brennan et al., 1995; Cronbach et al., 1972, p. 287) between hands-on and paper-and-pencil scores is

$$\frac{\sigma(_hp,{}_pp)}{\sqrt{\left[\sigma^2(_hp)\cdot\sigma^2(_pp)\right]}} \quad (4)$$

Just as analysis of variance can be used to obtain estimated components of covariance, multivariate analysis of variance provides a computational procedure for obtaining estimated components of variance and covariance. While analysis of variance provides scalar values for the sums of squares and mean squares, multivariate analysis of variance provides matrices of sums of squares and cross products and mean squares and cross products. Estimates of the variance components are obtained by setting the expected mean square equations equal to the observed mean squares and solving the set of simultaneous equations. Analogously, estimates of the components of covariance are obtained by setting the expected mean product equations equal to the observed mean products and solving the set of simultaneous equations.

Persons x tasks x raters design (using data for Occasion 1). The estimated variance and covariance components using these procedures are given in Tables 7a and 7b for the persons x tasks x raters design and in Table 8 for the persons x tasks x raters x occasions design. In Tables 7a and 7b and Table 8, the symbol Σ designates a matrix of variance and covariance components. Using the scores from the first

Table 7a

Occasion 1: Estimated Variance and Covariance Components and Disattenuated Correlations Between Hands-On and Paper-and-Pencil Scores for the *p x t x r* Design

	Hands-on	Paper-and-pencil
$\hat{\Sigma}_p$	.0450	.0407
	.0407	.0515
$\hat{\Sigma}_t$	0[a]	-.0001
	-.0001	0
$\hat{\Sigma}_r$	0	.0000
	.0000	.0000
$\hat{\Sigma}_{pt}$	.0323	.0026
	.0026	.0183
$\hat{\Sigma}_{pr}$	.0001	.0000
	.0000	.0000
$\hat{\Sigma}_{tr}$	.0000	.0000
	.0000	0
$\hat{\Sigma}_{ptr,e}$	.0005	-.0000
	-.0000	.0001
Observed *r*[b]	.69	
Disattenuated *r*	.85	

Note. Analysis uses data from Occasion 1 only. $\hat{\Sigma}$ denotes a matrix of estimated variance and covariance components: diagonal entries are estimated variance components; off-diagonal entries are estimated covariance components.

[a]Negative estimated variance component set equal to zero.

[b]Observed correlation between hands-on and paper-and-pencil total scores for Occasion 1.

occasion only, the elements of $\hat{\Sigma}_p$, the estimated variance-covariance component matrix for persons (*p*), are:

$$\hat{\sigma}(hp,pp) = .0407$$

$$\hat{\sigma}^2(hp) = .0515$$

$$\hat{\sigma}^2(pp) = .0450$$

Table 7b

Occasion 2: Estimated Variance and Covariance Components and Disattenuated Correlations Between Hands-On and Paper-and-Pencil Scores for the *p x t x r* Design

	Hands-on	Paper-and-pencil
$\hat{\Sigma}_p$	.0424	.0482
	.0482	.0582
$\hat{\Sigma}_t$	0[a]	-.0001
	-.0001	0
$\hat{\Sigma}_r$	0	-.0000
	-.0000	0
$\hat{\Sigma}_{pt}$	.0211	.0051
	.0051	.0207
$\hat{\Sigma}_{pr}$	.0000	.0000
	.0000	.0000
$\hat{\Sigma}_{tr}$	0	.0000
	.0000	.0000
$\hat{\Sigma}_{ptr,e}$	.0002	-.0000
	-.0000	.0001
Observed r[b]		
Disattenuated r	.97	

Note. Analysis uses data from Occasion 2 only. $\hat{\Sigma}$ denotes a matrix of estimated variance and covariance components: diagonal entries are estimated variance components; off-diagonal entries are estimated covariance components.

[a]Negative estimated variance component set equal to zero.

[b]Observed correlation between hands-on and paper-and-pencil total scores for Occasion 1.

The substantial estimated covariance component [$\hat{\sigma}(hp,pp)$] relative to the estimated variance components [$\hat{\sigma}^2(hp)$ and $\hat{\sigma}^2(pp)$] shows that students with high universe scores on the hands-on test tended to have high universe scores on the paper-and-pencil test. The disattenuated correlation using Equation 4 is .85, considerably higher than the observed correlation of .69. The disattenuated correlation is the estimated value of the correlation between hands-on and paper-and-pencil scores as the number of tasks and the number of raters each approaches infinity.

Nearly all of the estimated error covariance components in Table 7 are close to zero. For example, the small estimated covariance component for tasks ($\hat{\sigma}(ht,pt)$ = -.0001) shows that the high similarity of tasks used on the hands-on and paper-and-pencil tests does not contribute to the relationship between hands-on and paper-and-pencil mean scores. The estimated covariance component for the person x task interaction is somewhat larger, although still small ($\hat{\sigma}(hpt,ppt)$ = .0026), especially compared to the corresponding estimated variance components ($\hat{\sigma}^2(hpt)$ = .0323 and $\hat{\sigma}^2(ppt)$ = .0183). This shows that, although the relative standing of students differed across tasks within each test (the large variance components), those differences in rank orderings were not consistent across the two tests (the small covariance component).

Persons x tasks x raters x occasions design. The estimated components of variance and covariance for the persons x tasks x raters x occasions design are presented in Table 8. The disattenuated correlation, equal to .89, is the estimated value of the correlation between hands-on and paper-and-pencil scores as the number of tasks, raters, and occasions each approaches infinity. The disattenuated correlation is only somewhat higher than the high observed correlation between hands-on scores averaged over the two occasions and paper-and-pencil scores averaged over the two occasions, .83. The high observed correlation makes sense given the large effect of occasion shown in the univariate generalizability analyses (Table 5): Averaging over two occasions substantially reduces estimated error variation and thus raises the observed correlation between hands-on and paper-and-pencil scores.

While the disattenuated correlation between hands-on and paper-and-pencil scores in the persons x tasks x raters x occasions design is very similar to that for the persons x tasks x raters design in Table 7 (.89 vs. .85), substantial differences in the magnitudes of the components of covariance appear with the introduction of the occasion facet into the design. First, the estimated covariance component for the person x task interaction is larger in this design than in the previous design ($\hat{\sigma}(hpt,ppt)$ = .0065 vs. .0026) and is nearly as large as one of the estimated variance components in this design (hands-on scores: $\hat{\sigma}^2(hpt)$ = .0091). Averaging across scores for both occasions, then, there seems to be some consistency in the differences in rank orderings of students across tasks from the hands-on test to the paper-and-pencil test. For example, students who performed well on the first task on the

Table 8

Estimated Variance and Covariance Components and Disattenuated Correlations Between Hands-On and Paper-and-Pencil Scores for the *p x t x r x o* Design

	Hands-on	Paper-and-pencil
$\hat{\Sigma}_p$	.0375	.0381
	.0381	.0483
$\hat{\Sigma}_t$	0[a]	-.0001
	-.0001	.0000
$\hat{\Sigma}_r$	0	-.0000
	-.0000	.0000
$\hat{\Sigma}_o$	.0020	.0032
	.0032	.0049
$\hat{\Sigma}_{pt}$	.0091	.0065
	.0065	0
$\hat{\Sigma}_{pr}$	.0000	-.0000
	-.0000	0
$\hat{\Sigma}_{po}$	.0061	.0064
	.0064	.0066
$\hat{\Sigma}_{tr}$	.0000	.0000
	.0000	.0000
$\hat{\Sigma}_{to}$	0	.0000
	.0000	0
$\hat{\Sigma}_{ro}$	0	.0000
	.0000	0
$\hat{\Sigma}_{ptr}$	.0000	.0000
	.0000	0
$\hat{\Sigma}_{pto}$	.0176	-.0027
	-.0027	.0216
$\hat{\Sigma}_{pro}$	.0001	.0000
	.0000	.0000
$\hat{\Sigma}_{tro}$	0	-.0000
	-.0000	0
$\hat{\Sigma}_{ptro,\,e}$	.0003	-.0000
	-.0000	.0001

Table 8 (continued)

	Hands-on	Paper-and-pencil
Observed r[b]	.83	
Disattenuated r	.89	

Note. $\hat{\Sigma}$ denotes a matrix of estimated variance and covariance components: diagonal entries are estimated variance components; off-diagonal entries are estimated covariance components.

[a]Negative estimated variance component set equal to zero.

[b]Observed correlation between the average hands-on score across Occasions 1 and 2 and the average paper-and-pencil score across Occasions 1 and 2.

hands-on test (compared to their peers) but not well on the second task of the hands-on test tended to show a similar pattern of results on the paper-and-pencil test.

Two other estimated covariance components are non-negligible and they both involve the occasion facet. The first, the estimated covariance components for the person x occasion interaction, is fairly large ($\hat{\sigma}(hpo,ppo)$ = .0064) and is comparable to the corresponding estimated variance components ($\hat{\sigma}^2(po)$ = .0061 for hands-on, $\hat{\sigma}^2(po)$ = .0066 for paper-and-pencil). Differences in rank orderings of students from one occasion to the next were consistent across the hands-on and paper-and-pencil tests. For example, students who performed well on the hands-on test on the first occasion (compared to their peers) but not well on the second occasion showed a similar pattern of results on the paper-and-pencil test. The second non-negligible estimated covariance component is the one for occasions ($\hat{\sigma}(ho,po)$ = .0032), which is the same magnitude as the estimated variance components ($\hat{\sigma}^2(o)$ = .0020 for hands-on, $\hat{\sigma}^2(o)$ = .0049 for paper-and-pencil). This result shows that mean differences in performance across occasions were consistent across the hands-on and paper-and-pencil tests (Table 1 showed that, on both tests, scores increased from the first to the second occasion).

The non-negligible role of occasion in the estimated covariance components shows that administering the tests on the same occasions contributed somewhat to the relationship between hands-on and paper-and-pencil scores. This suggests that observed correlations between tests administered on the same occasion should be

somewhat higher than observed correlations between tests administered on different occasions. The observed correlations in this study conform to this prediction. The correlations between the hands-on and paper-and-pencil tests were .69 for the first occasion and .84 for the second occasion, averaging to .77. The correlations between hands-on and paper-and-pencil tests administered on different occasions were .70 and .66, averaging to .68. Although the difference between average correlations is not large, it is in the expected direction.

Disattenuated correlations between hands-on and paper-and-pencil scores. As noted above, the disattenuated correlation between hands-on and paper-and-pencil scores is very similar whether the design used is persons x tasks x raters (.85) or persons x tasks x raters x occasions (.89). The similarity of the disattenuated correlations shows that the universe-score correlation between hands-on and paper-and-pencil scores is similar whether the focus is on the first occasion of testing or the average across all occasions. So the introduction of occasion as a source of error variability in this study has little impact on the disattenuated correlation.

Although the disattenuated correlations estimated from the two designs were very similar in this study, this need not be the case. As described earlier, using a persons x tasks x raters x occasions design, the effects involving occasion can be estimated separately. In the persons x tasks x raters design, however, occasions is a hidden facet. Although effects involving occasion cannot be estimated separately, they are still present, although confounded with the other effects in the design. Consequently, each estimated variance component is really a combination of two effects (e.g., the variance components for *p* and *po*).

A similar confounding applies to covariances in the multivariate design. The covariances involving occasion are still present; they just cannot be estimated separately. For example, just as what is interpreted as universe-score variance in the persons x tasks x raters design is really $\sigma^2(p) + \sigma^2(po)$, what is interpreted as universe-score covariance is really $\sigma(h_p,p_p) + \sigma(h_{po},p_{po})$. The covariance component for the person x occasion interaction—$\sigma(h_{po},p_{po})$—will contribute to the numerator of the disattenuated correlation coefficient (see Equation 4), and the variance component for the person x occasion interaction for each test format—$\sigma^2(h_{po})$ and $\sigma^2(p_{po})$—will contribute to the denominator of the disattenuated coefficient.

The disattenuated correlation estimated using variance and covariance components from the persons x tasks x raters design may be larger or smaller than

the disattenuated correlation estimated from the persons x tasks x raters x occasions design depending on the relative magnitudes of the variance and covariance components for the person x occasion interaction. In the present study, the magnitudes of the variance and covariance components for the person x occasion interaction were such that the overall ratio (Equation 4) remained much the same whether the *po* effects were included or not. In a study in which the *po* variance components are substantial and the *po* covariance component is negligible, the disattenuated correlation estimated using a persons x tasks x raters design will be smaller than the disattenuated correlation estimated using a persons x tasks x raters x occasions design.

The high disattenuated correlations in the present study show that students with high universe scores on the hands-on test also tend to have high universe scores on the paper-and-pencil test. Because interest is normally in generalizing over occasions of testing, as well as over tasks and raters, the relevant universe of generalization includes multiple tasks, raters, and occasions. The disattenuated correlation of .89, then, suggests that paper-and-pencil scores may be considered interchangeable with hands-on scores if sufficiently large numbers of tasks, raters, and occasions are used to reduce error variability to near zero. Whether error-free scores are obtainable in practice, however, is questionable. The results of univariate analyses presented earlier (Table 6) showed that, when the decision maker uses one occasion of testing, a large number of tasks would be needed to obtain estimated generalizability coefficients as high as .80: four tasks for the paper-and-pencil test and eight tasks for the hands-on test. Even with an extremely large number of tasks, say 100, estimated generalizability coefficients for one occasion would reach only .85 for hands-on scores and .88 for paper-and-pencil scores. While these coefficients are fairly high, the scores cannot be considered error-free. For one occasion of testing, then, it is probably impossible to obtain hands-on and paper-and-pencil scores that would be close to the disattenuated correlation of .89 or would correlate highly enough to be considered interchangeable.

Discussion

This study set out to investigate the importance of occasion as a hidden source of error variance in (a) estimates of the dependability (generalizability) of science assessment scores and (b) the interchangeability of science test formats. The findings of the univariate generalizability analyses in the current study confirm those found

by McBee and Barnes (1998) and Shavelson et al. (1999). As in the previous studies, analyses of the scores in the present study showed that, when occasion was not considered as an explicit facet, task-sampling was the major source of variation. For both the hands-on and paper-and-pencil tests, the person x task interaction effect was quite large, showing that the relative standing of examinees was not consistent from one task to the other. Differences between raters, on the other hand, were very small. Averaging over two tasks (each taking 20-25 minutes), the levels of estimated generalizability were .73 and .80 for hands-on scores for each occasion, and .85 for paper-and-pencil scores for each occasion.

Adding occasion to the design as a source of error variation altered the interpretation about the substantial sources of error in the measurement. As in the previous studies, the largest source of error variation was due to a combination of task-sampling and occasion-sampling: the person x task x occasion interaction. This effect shows the volatility of student performance across the combinations of tasks and occasions. In the present study, the person x occasion interaction effect was also substantial, showing that students did not maintain the same relative standing across occasions. The importance of task-sampling by itself was reduced when occasion was included in the design: the percentage of the total variation due to the person x task interaction effect was reduced by half in the analysis of hands-on scores and was reduced to zero in analysis of paper-and-pencil scores. Taking occasion explicitly into account also changed the estimated level of generalizability of the test scores: the estimated generalizability coefficients for a single occasion were reduced to .66 (hands-on test) and .73 (paper-and-pencil test). These results show that, as Cronbach et al. (1997) described, ignoring occasion of measurement may produce inflated estimates of generalizability. In this study, the magnitude of the inflation is practically significant for a two-task test (.73 vs. .66 for hands-on scores; .85 vs. .73 for paper-and-pencil scores).

Adding occasion as a source of variance in the multivariate generalizability analyses also influenced the interpretation of the observed correlation between hands-on and paper-and-pencil scores. The observed correlations between hands-on and paper-and-pencil scores were fairly high (.69 on the first occasion, .84 on the second occasion, .83 for the average score across the two occasions). The multivariate analyses for the two designs yielded somewhat different information about the sources of error covariation contributing to the observed correlation. The multivariate analyses of scores from the first occasion (persons x tasks x raters

design) showed that neither tasks nor raters were sources of error covariation among hands-on and paper-and-pencil scores. The multivariate generalizability analyses of scores from both occasions (persons x tasks x raters x occasions design), in contrast, showed that both tasks and occasions contributed to the high observed correlations. The non-negligible person x task covariance component showed that differences in relative standings of examinees across tasks were consistent across the hands-on and paper-and-pencil tests, suggesting that the near-identical structure of the tasks on the two tests contributed to the observed correlation between scores on the two tests. The non-negligible person x occasion covariance component showed that taking the tests on the same occasion contributed to the observed correlation. This result suggests that administering the tests on the same occasion may produce an inflated picture of the relationship between the two testing methods. This is supported by the somewhat higher observed correlations between tests administered on the same occasion than the observed correlations between tests administered on different occasions.

Estimates of the disattenuated correlations from the multivariate generalizability analyses were high and similar for both designs—.85 for the persons x tasks x raters design using the data from Occasion 1 and .89 for the persons x tasks x raters x occasions design. Although the two designs produced similar estimates of correlation between universe scores on the two types of tests in the current study, they need not be similar. As described earlier, because the person effect is confounded with the person x occasion interaction effect in a design that does not include occasions as an explicit facet (here, a persons x tasks x raters design), and this confounding does not occur in a design that includes occasions as an explicit facet (here, a persons x tasks x raters x occasions design), the estimated disattenuated correlations from the two designs will differ according to the relative magnitude of the variance and covariance components for the person x occasion interaction. More research is needed to be able to predict what magnitudes of the *po* effects (both variance and covariance components) are likely to arise in practice. In the present study, for example, the estimated variance components for the *po* effects were substantial, whereas in the McBee and Barnes (1998) and Shavelson et al. (1999) studies, they were negligible.

Both the univariate and multivariate generalizability analyses provide evidence of the volatility of student performance over occasions in the present study. Shavelson et al. (1999) suggested that the volatility of performance may be due to

students' partial knowledge of the science they apply to performance assessments. Evidence supporting this hypothesis comes from Troper's (1998) examination of the changes over time in students' mental models about electrical resistance. Troper analyzed changes in mental models as reflected on the paper-and-pencil test responses of the students in the present project who worked in collaborative small groups on the hands-on test between the two administrations of the paper-and-pencil test. On both administrations of the paper-and-pencil test, the vast majority of students showed incomplete and often incorrect understanding of the concept of resistance, its causes, its effect on brightness of a circuit, and its relation to related concepts such as current. Furthermore, students often seemed to be applying different conceptions across different items of the same test administration (for example, stating on some items that resistors "use up" or consume current on some items, and stating on other items that resistors block or impede current; or giving different causes of resistance—bulbs, graphite resistors, wires, batteries—on different items). Moreover, a substantial portion of Troper's sample (about 20%) showed a decrease in the sophistication or coherence of their mental models about resistance from the first administration of the paper-and-pencil test to the second. Because the sample analyzed in the present study was carefully matched to the sample in Troper's study, it is highly likely that students in the present study show the same range of misconceptions and incomplete knowledge about electricity. And, as in Troper's study, a substantial proportion of the students in the present sample showed a decrease in their scores from the first administration of the test to the second (32% of the sample showed a decrease on the hands-on test; 23% showed a decrease on the paper-and-pencil test).

That students had partial knowledge was not surprising given the nature of instruction in this study. Teachers spent only about three weeks teaching the concepts of voltage, resistance, current, electric circuitry, and (sometimes) Ohm's Law. Furthermore, in an analysis of students' opportunity to learn based on teachers' instructional materials, lesson plans, and interview responses about their instruction during the unit on electricity, Wang (1997) found that students in many classes had little opportunity to learn the concepts well. Not all teachers taught all of the concepts, some did not teach the relationships among the concepts, and teachers estimated that, on the average, only about half of the students would be able to do well on the tests. Moreover, no class had available all of the equipment used on the test (resistors, batteries, bulbs, alligator clips and wires), and some classes had no

equipment at all. While most classes had textbooks, in most schools students were not allowed to take textbooks home. It is highly unlikely that many students in this study had enough time and opportunity to formulate correct conceptions and to practice applying their knowledge in order for it to be automated. Fitts and Posner (1967), Anderson (1983), and others have suggested that, among three stages of learning—cognitive or conscious, associative, and automated—only the automated stage is likely to lead to consistent performance.

In conclusion, the results of this study show that ignoring occasion as a source of variation can seriously overestimate the dependability of achievement test scores, whether hands-on performance tests or paper-and-pencil tests; may lead to misleading conclusions regarding other sources of error in the measurement; and may lead to misinterpretations about the sources of error that contribute to the correlations between scores on different types of tests.

References

Anderson, J. A. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Bennett, R., & Ward, W. (Eds.). (1993). *Construction versus choice in cognitive measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Brennan, R. L. (1992). *Elements of generalizability theory* (rev. ed.). Iowa City, IA: American College Testing.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys Listening and Writing Tests. *Educational and Psychological Measurement, 55,* 157-176.

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

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373-399.

Fitts, P. M., & Posner, M. I. (1967). *Human performance mastery*. Monterey, CA: Brooks / Cole.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7,* 323-342.

Gao, X., Shavelson, R. J., Brennan, R. L., & Baxter, G. (1996, April). *A multivariate generalizability theory approach to convergent validity of performance-based assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

McBee, M. M., & Barnes, L. L. B. (1998). The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education, 11,* 179-194.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics.* Reston, VA: Author.

National Research Council. (1996). *National science education standards.* Washington, DC: National Academy Press.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30,* 41-53.

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

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36,* 61-71.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory—1973-1980. *British Journal of Mathematical and Statistical Psychology, 34,* 133-166.

Sugrue, B., Webb, N., & Schlackman, J. (1998, April). *The interchangeability of assessment methods in science.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego.

Troper, J. D. (1997). *The effects of small-group discussion on students' learning and transfer of ideas in science.* Unpublished doctoral dissertation, University of California, Los Angeles.

Wang, J. (1996). *A multilevel modeling study of opportunity to learn (OTL) and science achievement*, Unpublished doctoral dissertation, University of California, Los Angeles.

Webb, N. M., Shavelson, R. J., & Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Fyans (Ed.), *Generalizability theory: Inferences and practical applications* (pp. 67-81). San Francisco, CA: Jossey-Bass.

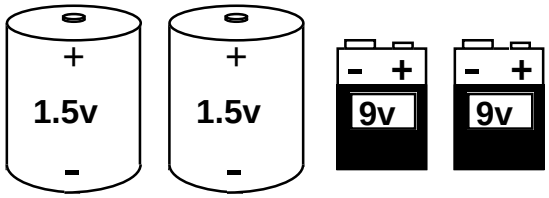
Appendix A

Example Test Items From Paper-and-Pencil Test

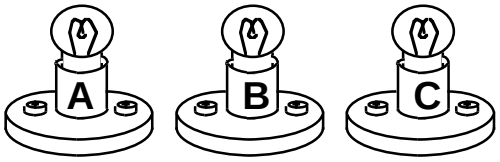
1. (a) Use the items drawn below (batteries and bulbs) to draw **two circuits** in the boxes labeled Circuit A and Circuit B. **Follow these rules:**

- **Bulb A should be in Circuit A. Bulb B should be in Circuit B.**
- **Bulb A should be brighter than Bulb B**.
- **There should be one 9-volt battery in each circuit.**
- **You must draw the wires needed to connect up the items in each circuit.**
- **Use all of the items but do not use any item more than once.** For example, if you put Bulb C in Circuit A, you cannot also put it in Circuit B.

Batteries



Bulbs



Draw the circuits in these boxes:

Circuit A (brighter)	Circuit B (dimmer)

1. (b) Why will Bulb A in Circuit A be brighter than Bulb B in Circuit B? (Try to use scientific terms in your answer.)

_____

_____

_____

1. (c) Which of the two circuits you drew has the **highest voltage**?

Circle one: **CIRCUIT A** **CIRCUIT B** **BOTH CIRCUITS HAVE THE SAME VOLTAGE**

Why? (Try to use scientific terms in your answer.)

_____

_____

_____

1. (d) Which of the two circuits you drew has the **highest resistance**?

Circle one: **CIRCUIT A** **CIRCUIT B** **BOTH CIRCUITS HAVE THE SAME RESISTANCE**

Why? (Try to use scientific terms in your answer.)

_____

_____

_____

1. (e) Which of the two circuits you drew has the **highest current**?

Circle one: **CIRCUIT A** **CIRCUIT B** **BOTH CIRCUITS HAVE THE SAME CURRENT**

Why? (Try to use scientific terms in your answer.)

_____

_____

_____