The Interchangeability of Assessment Methods in Science

CSE Technical Report 474

Brenda Sugrue
The University of Iowa

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

August 1998

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/University of California, Los Angeles

The work reported herein was supported 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 Officer of Educational Research and Improvement, or the U.S. Department of Education.

THE INTERCHANGEABILITY OF ASSESSMENT METHODS IN SCIENCE*

Brenda Sugrue
The University of Iowa

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

Abstract

This article describes a study that investigated the interchangeability of four different assessment methods for measuring middle-school students' understanding of science concepts. The four methods compared were hands-on tasks with associated multiple-choice and written justification items, written analogues of the hands-on tasks, and two types of multiple-choice items that were not related to hands-on or written analogue tasks. Some students took the hands-on test before the written test, and some students took the written test before the hands-on test. Observed and disattenuated correlations were examined. Multivariate generalizability analysis was used to obtain disattenuated correlations. The results indicate that hands-on and written analogue tests are not interchangeable, but multiple-choice and written justification items linked to hands-on and written analogues could be considered interchangeable if correlations between .76 and .96 are an acceptable criterion for interchangeability. In addition, a number of interesting order effects were found.

Diversification of methods of assessment in both classroom and large-scale testing has generated renewed interest in the comparability or interchangeability of methods (see, for example, Bennett & Ward, 1993). It is generally assumed that more "authentic" and costly methods of assessment, such as hands-on performance tasks in science, yield more valid estimates of student knowledge than do more efficient methods, such as paper-and-pencil multiple-choice items, although a number of authors (for example, Royer, Cisero, & Carlo, 1993, and Schmidt & Bjork, 1992) suggest that assessment and practice activities can be cognitively authentic—that is, can elicit the kinds of cognitive processing characteristic of expertise in a domain—without being contextually authentic.

* Paper presented at the 1998 annual meeting of the National Council on Measurement in Education, San Diego, Thursday, April 16.

Recent research also indicates that the cognitive demands of assessments are not necessarily a function of their format. Multiple-choice items can induce complex thinking in a domain (Hamilton, Nussbaum, & Snow, 1995); and performance tasks can fail to induce higher level processing if they give very detailed, step-by-step instructions. Even if tasks are challenging, scoring schemes can be insensitive to differences among students in deeper levels of understanding in a domain (Baxter, Glaser, & Raghavan, 1993). To further complicate the situation, overall scores on different methods, such as multiple-choice and open-ended questions, can be comparable, but performance can vary considerably at the individual item level (Bridgeman, 1992). Given the variety of findings from studies that have compared methods of assessment, it would be difficult to predict which methods might and which methods might not be interchangeable for a particular body of knowledge or group of students. There is clearly a need for further research on this issue.

Only one study has directly addressed the issue of assessment method interchangeability in the domain of science (Baxter & Shavelson, 1994). Baxter and Shavelson found that scores on notebook surrogates of science performance assessments were comparable to scores based on real-time observations of student performance on hands-on tasks (mean scores were similar, and correlations were between .75 and .84). Scores generated by other methods of assessment (computer simulation, multiple-choice items, and short-answer items) were not comparable to scores based on hands-on tasks; mean scores were similar, but correlations ranged from only .28 to .53.

Although scores based on observation of hands-on performance seem to be the ideal or "benchmark" assessment in science, scores generated by performance assessments have themselves proved to be unstable with large variability across tasks (Lane, Liu, Ankenmann, & Stone, 1996; Shavelson, Baxter, & Gao, 1993). This instability has cast doubt on the feasibility of using hands-on tasks in large-scale or high-stakes testing situations. The solution often suggested, based on person-by-task generalizability studies, is that at least ten tasks are needed to arrive at stable estimates of performance in whatever domain is being sampled (Dunbar, Koretz, & Hoover, 1991). Others suggest that greater comparability can be achieved by using templates to standardize the structure of science tasks or by standardizing the dimensions of scoring rubrics (e.g., Solano-Flores & Shavelson, 1997).

Nichols and Sugrue (1997) have suggested that greater comparability might result from adopting a more "construct-centered" approach to assessment. In a construct-centered approach to assessment, the cognitive aspects of student knowledge to be assessed are clearly defined, and both assessment tasks and scoring rubrics are aligned with the definition of the constructs. For example, if the cognitive constructs being measured were knowledge of concepts and the relationships among concepts in a particular domain of science, then a set of items would be constructed that would diagnose which concepts and which relationships a student understood. Alternatively, an extended task could be constructed and separate elements of a response would generate scores linked to specific concepts and relationships. If the flexibility of a student's knowledge base was the construct of interest, then a set of tasks with ever more varied and complex situations could be constructed, although all of the tasks would require application of the same basic knowledge.

The opposite of a construct-centered approach might be called a content-centered approach whereby test specifications consist primarily of numbers of items per content area. Even in cases where content-by-process matrices are used to construct test specifications, the process categories are usually very general (e.g. conceptual understanding, procedural knowledge and problem solving), and the attributes of items that would fit the content and process categories are not clearly defined. Thus, it is difficult to generate isomorphic items or to predict the relative difficulty of items.

The study reported here used a construct-centered approach to create a set of assessments that would measure students' understanding of a few specific concepts that are normally part of the eighth-grade science curriculum. In this study, the broad construct to be measured was student understanding of the relationships between the abstract concepts of voltage, resistance, and current, and the concrete components of electric circuits (for example, batteries and bulbs). It was reasoned that students who understood these relationships would be able to do the following kinds of activities:

- make circuits to match particular specifications;
- draw circuits to match particular specifications;
- identify circuits that have higher voltage, resistance, and current;

- explain why circuits have higher voltage, resistance, and current;
- predict what would happen to the voltage, resistance and current in circuits if changes were made to the components of the circuits.

To measure the ability constructs listed above, four types of assessments were created:

1. hands-on tasks with related diagram-drawing, multiple-choice and written justification items;
2. paper-and-pencil versions of the hands-on tasks; these analogous tasks were similar to the hands-on tasks in all respects with the exception of the opportunity to manipulate real circuit components;
3. a set of multiple-choice items that asked students to identify circuits with the highest voltage, resistance and current;
4. a set of multiple-choice items that asked students to predict how the voltage, resistance, and current in a circuit would change if certain changes were made to its components.

With this set of items it was possible to compare the interchangeability of four different assessment methods. In addition, we were able to (a) compare scores on multiple-choice and open-ended justification items in the context of hands-on and analogous tasks, and (b) estimate the contribution of hands-on manipulation of equipment to performance. The results have implications for balancing assessment authenticity and efficiency.

Method

Study Design

662 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. At the beginning of the study, students were administered tests of vocabulary, verbal reasoning, and nonverbal reasoning; these were used to control for differences in general ability. Then all teachers conducted a three-week unit on electricity and electric circuits in their classrooms. 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 the hands-on and written

tests. The order of tests was counterbalanced so that half of the students took the hands-on test on the first day and the other half took the written test on the first day.

Instruments

Tests of general cognitive abilities. Two measures of general ability were obtained, one a measure of nonverbal reasoning (Raven's Progressive Matrices) and one a measure of verbal reasoning (The New Jersey Test of Verbal Reasoning). A vocabulary test (Ekstrom, French, Harman, & Dermen, 1976) was also administered.

Science assessments. The hands-on test consisted of two tasks. Each task required students to assemble batteries, bulbs, wires, and (in one task) resistors into two electric circuits so that one circuit was brighter (or dimmer) than the other, draw their circuits, and then answer three questions about their circuits. Each question had two items measuring knowledge of one the three concepts (voltage, resistance, or current): a multiple-choice item (e.g., "Which circuit had higher voltage?") and a written justification item ("Why?"). The two tasks on the hands-on test differed only in the number and type of circuit components provided; the equipment for the first task consisted of two 1.5 volt batteries, two 9 volt batteries, three bulbs and six wires; the equipment for the second task consisted of two 9 volt batteries, three graphite bars (resistors), two bulbs, and six wires. Nearly all analyses used only the scores on the multiple-choice and justification responses related to the hands-on and written analogues. Scores assigned to the circuits students made or drew were used in some follow-up exploratory analyses.

The paper-and-pencil written test had three parts. One part, analogous to the hands-on test (called the written analogue), asked students the same multiple-choice and open-ended justification questions but showed them pictures of the equipment instead of giving them equipment. The second set of items on the written test asked students to identify the circuit that had the highest voltage, highest resistance, or highest current from groups of four circuits. The third set of items on the written test used a multiple-choice format to probe students' ability to predict what would happen to the voltage, resistance, and current in a circuit if changes were made to its components. Samples of all items are included in the Appendix.

Scoring

Multiple-choice items were scored dichotomously (right/wrong). The accuracy of multiple-choice and justification items related to the hands-on and analogue tasks was judged with respect to the particular circuits each student drew. For example, if a student drew the following circuits for Task 1 (see Figure 1), where circuit A has one 9-volt battery, one 1.5-volt battery and one bulbs, and circuit B has one 9-volt battery, one 1.5-volt battery and two bulbs), then to be correct, the student would have to select the responses "C (same)," "A," and "A" to the questions "Which circuit has the highest voltage?", "Which circuit has the highest resistance?" and "Which circuit has the highest current?"

Scores for written justifications on the hands-on and analogue tasks were based on the extent to which students referred to concrete components of the circuits drawn and the abstract concepts they represented. Justifications for the relative voltage in the circuits a student made or drew were scored on a 4-point scale. A score of 0 was assigned if a student gave an irrelevant answer or displayed confusion over cause and effect, such as "the voltage is higher because it is brighter." A score of 1 was assigned if the student mentioned batteries (but not the relative number) as the source of different voltage in the two circuits. A score of 2 was assigned if a student mentioned the relative number of batteries in each circuit. A score of 3 was assigned if a student mentioned the relative number of batteries and also referred to the relative power or voltage generated

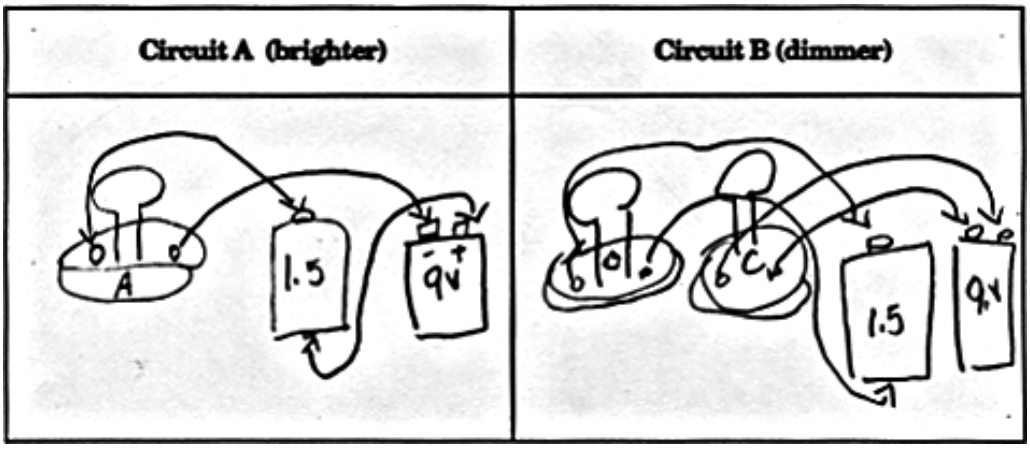


Figure 1. Circuits drawn for hands-on Task 1.

by the batteries. In the example given earlier, the student's justification for voltage would need to include a reference to the similar number of volts in the batteries in circuits A and B in order to obtain the highest score.

Justifications for the relative resistance in the circuits students made or drew were scored on a 6-point scale. Again, 0 was assigned for irrelevant answers or answers confusing cause and effect. A score of 1 was assigned if the student referred to a difference in the number of "things" in the circuits. A score of 2 was assigned if a student was more specific and mentioned either wires or the distance the electrons had to travel. A score of 3 was assigned if the student mentioned graphite or bulbs as the cause of differences in resistance between the circuits. A score of 4 was assigned if the student described the number of items causing the difference in resistance. A score of 5 was assigned if a student referred to the exact numbers and types of items causing the difference in resistance. In the example given earlier, the student's justification for resistance would need to include a reference to the greater number of bulbs in Circuit B than in Circuit A in order to obtain the highest score.

Justifications for the relative current in the circuits a student made or drew were coded along three dimensions (reference to voltage, reference to resistance, and reference to brightness) and a composite score was calculated based on a combination of these three codes. The voltage and resistance dimensions were coded on a 3-point scale where 2 meant that a student referred to the more abstract concept of voltage or resistance, 1 meant that the student referred to the more concrete concepts of batteries, bulbs or graphite, and 0 meant that the student did not mention voltage or resistance (either abstract or concrete) as the cause of current differences between the two circuits. The third dimension coded (mention of brightness as a cause or effect of current) was coded dichotomously. The composite variable representing performance on current justifications was formed by summing the codes for the voltage and resistance dimensions, and giving half a point to a student who mentioned brightness but scored 0 on the voltage and resistance dimensions. In the example given earlier, the resistance dimension of a student's justification for current would need to include reference to the greater number of bulbs in Circuit A in order to obtain the highest score.

Scores on the justification items were converted to scales from 0 to 1 by dividing the score by the maximum possible number of points. This resulted in

all score variables being on a 0 to 1 scale for analysis. A number of aggregate scores were created. Total scores on each of the four types of assessment (hands-on, written analogues, identification, and prediction) were generated by computing the mean of all items in the assessment type. The hands-on total score was the mean of the six multiple-choice and six justification items across the two hands-on tasks. The written analogue total score was the mean of the multiple-choice and justification scores on the two analogous tasks. The identification total score was the mean of scores on the three pairs of items that measured knowledge of voltage, resistance, and current. The prediction total score was the mean of scores on the three sets of six items that measured knowledge of voltage, resistance, and current.

Diagrams of circuits that students drew were scored according to the degree to which they followed instructions (e.g., using every piece of equipment once and only once, including a 9-volt battery in each circuit). The following ordered scale was used for the first task: 9 (used all pieces of equipment as instructed), 8 (used all equipment but put both 9-volt batteries in one circuit), 7 (omitted one battery), 6 (omitted two batteries), 5 (omitted one bulb), 4 (put both 9-volt batteries in one circuit and omitted one bulb), 3 (omitted one bulb and one battery), 2 (omitted two batteries and one bulb), and 1 (other, such as using the same piece of equipment in both circuits, constructing one instead of two circuits, constructing the same circuit twice). The following ordered scale was used for the second task: 4 (used all pieces of equipment as instructed), 3 (omitted one graphite resistor), 2 (omitted two graphite resistors), and 1 (other, such as using the same piece of equipment in both circuits, constructing one instead of two circuits, constructing the same circuit twice).

Analysis

Three types of analysis were performed. First, observed correlations were used to examine the consistency of student performance across items of different types (multiple-choice vs. justification) within the hands-on and analogue tasks, and across tests with different formats or requirements (hands-on, written analogue, written prediction, written identification). Second, univariate generalizability analyses were conducted to examine the reliability of performance across items on each of the four assessment methods (hands-on, written analogue, written prediction, written identification). Third, multivariate generalizability was used to examine universe-score correlations disattenuated

for measurement error. Because of significant order effects (depending on which test the student completed first, described in detail below), all analyses were conducted separately for each order of hands-on and written tests.

Samples for analysis. Because of missing data, the samples used for analysis were considerably lower than the 662 students who participated in the study. Only 344 students were present for the administration of both the hands-on and written analogue tests and attempted both tasks on each test; therefore, only the data for these 344 students were used in analyses of performance on the hands-on and written analogue tasks. Eleven of these 344 students did not respond to any of the written identification items; therefore, data for 333 students were used for analyses that involved scores on identification items. Twenty of the 344 students did not respond to any of the written prediction items; therefore, data from 324 students were used for analyses that involved scores on prediction items.

Results

Results are presented in the following sequence:

- order effects;
- interchangeability of multiple-choice and justification item types;
- interchangeability of hands-on and written analogue tests;
- interchangeability of hands-on and written non-analogous items.

Order Effects

Order effects on mean scores. Table 1 presents descriptive statistics for the four assessment methods for the entire sample and for students who took each test on the first and second days of test administration. In addition to mean scores for each method, mean scores on parts of the hands-on and written analogues are given (each task and each item type), along with *t*-tests of the differences between the scores of students who took the particular test on the first day and the scores of students who took the test on the second day. In nearly all cases, performance on the hands-on and written analogue tests was significantly higher when the test was taken on the second day than when it was taken on the first day; the only exception was for scores on Task 1 of the written analogue.

Table 1

Order Effects on Mean Scores

Score	Whole sample		Took test on Day 1		Took test on Day 2		Independent samples *t*-test
	Mean	*SD*	Mean	*SD*	Mean	*SD*	
Hands-on total	0.50	0.26	0.44	0.24	0.54	0.26	3.75**
Written total	0.47	0.26	0.43	0.24	0.51	0.27	3.02**
Written identification	0.45	0.3	0.44	0.29	0.46	0.31	0.41
Written prediction	0.51	0.22	0.51	0.18	0.51	0.25	0.25
Hands-on Task 1	0.49	0.28	0.44	0.27	0.53	0.29	2.84**
Hands-on Task 2	0.5	0.3	0.44	0.3	0.56	0.29	3.76**
Written Task 1	0.5	0.27	0.48	0.25	0.52	0.29	1.66
Written Task 2	0.44	0.31	0.38	0.3	0.5	0.3	3.63**
Hands-on multiple-choice	0.58	0.3	0.52	0.3	0.63	0.3	3.26**
Hands-on justification	0.41	0.25	0.36	0.24	0.46	0.26	3.75**
Written multiple-choice	0.53	0.31	0.5	0.3	0.57	0.33	2.13*
Written justification	0.41	0.26	0.36	0.23	0.46	0.28	3.44**

Note. Day 1 Hands-on $n = 156$; Day 1 Written $n = 186$; Day 1 Written Identification $n = 156$; Day 1 Written Prediction $n = 155$; Day 2 Written Identification $n = 177$; Day 2 Written Prediction $n = 169$.

* = significant at .05 level. ** = significant at .01 level or beyond.

Performance on both multiple-choice and justification items was significantly higher on the test that was taken on the second day. Although the pattern of results was the same for written identification and prediction scores, the differences between identification or prediction scores of students who took the test on the first day and the scores of students who took the test on the second day were not statistically significant. It should be noted that there were no differences between the vocabulary, verbal reasoning, and nonverbal reasoning scores of students who took the tests in different orders.

Table 2 shows the results of paired *t*-tests of improvement in performance from Day 1 to Day 2 on comparable parts of the hands-on and written analogue tests. The improvement in scores from Day 1 to Day 2 was statistically significant regardless of which test (hands-on or written) was taken first.

To explore the nature of improvement from the first day to the second day, the performance of students who drew the same circuits on their papers on the two days was examined first. Because these students presumably constructed and

Table 2

Improvement From Day 1 to Day 2 on Analogous Sets of Items (Paired *t* Values)

Items/Scores	Hands-on Day 1 (n = 158)	Written Day 1 (n = 186)
Total	4.59**	6.78**
Task 1	4.04**	2.46*
Task 2	2.82**	8.34**
Multiple choice	2.00*	5.81**
Justification	6.30**	5.93**

* = significant at .05 level. ** = significant at .01 level or beyond.

drew the same circuits on both days, they would be most likely to give the same responses to the multiple-choice and justification items and, therefore, least likely to show improvement. Yet even these students showed improvement in their scores from Day 1 to Day 2. At the item level, improvement from Day 1 to Day 2 was often statistically significant. Students who took the hands-on test first showed statistically significant gains on one out of six multiple-choice items and four out of six justification items. Students who took the written test first showed statistically significant gains on four multiple-choice items and three justification items.

Inspection of these students' justifications showed that they often corrected a misconception about a concept that they had given on the test on Day 1. For items concerning voltage, for example, many students gave justifications for why one circuit had higher voltage than another on Day 1 in terms of bulbs, resistance, electricity, power, current, brightness, or energy. On Day 2, they correctly attributed the difference in voltage to the number and voltage of each of the batteries (e.g., a 1.5-volt and a 9-volt battery). For items concerning resistance, students on Day 1 often attributed the relative resistance of the two circuits to the batteries (or voltage of the batteries), power, or energy. On the second day, many of these students gave the correct justification in terms of the number of bulbs or graphite resistors.

The improvement of students who constructed different circuits on Days 1 and 2 was examined next. These students also showed statistically significant

improvement on many items. Students who took the hands-on test first showed statistically significant gains on two out of six multiple-choice items and three out of six justification items. Students who took the written test first showed statistically significant gains on three multiple-choice items and three justification items. Students who constructed different circuits on the two days showed similar corrections of misconceptions in their justifications as did students who constructed the same circuits on both days (e.g., incorrectly attributing resistance to batteries or power on the first day and correctly attributing resistance to bulbs or graphite resistors on the second day).

In addition to improving their justifications of why one circuit had higher voltage, resistance, or current than the other circuit, many students who constructed different circuits improved the quality of their circuits from Day 1 to Day 2. On the first task, 50% of students showed an increase in their circuit scores from Day 1 to Day 2, 26% showed no change in circuit scores (even though their exact circuits changed), and 24% showed a decrease in their circuit scores. On the second task, 55% of students showed an increase in their circuit scores from Day 1 to Day 2, 28% showed no change in circuit scores (even though their exact circuits changed), and 17% showed a decrease in their circuit scores. The percentages were very similar whether students took the hands-on test on Day 1 or the written analogue test on Day 1.

Order effects on correlations. Not only were the mean scores higher on the test taken on the second day, but the correlations among subscores within a test were also higher on the test taken on the second day. Table 3 shows correlations between subscores within the hands-on test and within the written analogue test.

Table 3

Order Effects on Correlations Among Parts of Hands-on and Written Analogues

Parts of test	Whole sample	Took test Day 1	Took test Day 2	*z* test
Hands-on Task 1 and hands-on Task 2	0.57	0.48	0.62	1.85
Written Task 1 and written Task 2	0.6	0.48	0.71	3.35*
Hands-on multiple choice and hands-on justification	0.73	0.65	0.77	2.25*
Written multiple choice and written justification	0.62	0.6	0.62	0.29

* = significant at .05 level.

Correlations between task scores, and between item type scores (multiple-choice and justification) were higher when either test was taken on the second day. Correlations between subscores within tests taken on the first day ranged from .48 to .65, whereas correlations for tests taken on the second day ranged from .62 to .77. Two of the four comparisons between first- and second-day correlations were statistically significant: correlations between tasks on the written analogue test, and correlations between multiple-choice and justification scores on the hands-on test.

Correlations between the written analogue and other parts of the written test (identification and prediction items) were also higher when the written test was taken on the second day. Table 4 shows that correlations on Day 1 ranged from .43 to .51, but correlations on Day 2 ranged from .52 to .68. Two of the three comparisons between first- and second-day correlations shown in Table 4 were statistically significant: correlations between written analogue scores and scores on prediction items, and correlations between scores on identification and prediction items.

Table 5 shows the correlations between written and hands-on tests. Correlations between hands-on and written analogue scores were higher when the hands-on test was taken on the first day than when the written test was taken on the first day. Correlations between written analogue and hands-on scores ranged from .57 to .72 when the hands-on test was taken first, but ranged from .46 to .60 when the written test was taken first. Three of the five comparisons of these first-day and second-day correlations were statistically significant: correlation between total scores on hands-on and written analogues, correlation between multiple-choice items (but not justification items) on both versions of

Table 4

Order Effects on Correlations Among Parts of Written Test

Parts of test	Whole sample	Took test Day 1	Took test Day 2	*z* test
Written total and written identification	0.5	0.48	0.52	0.49
Written total and written prediction	0.54	0.43	0.65	2.89*
Written identification and written prediction	0.59	0.51	0.68	2.44*

* = significant at .05 level.

Table 5

Order Effects on Correlations Between Hands-on and Written Tests

Tests	Whole sample	Hands-on Day 1	Written Day 1	*z* test
Hands-on and written analogues				
Hands-on total and written total	0.59	0.71	0.58	2.06*
Hands-on Task 1 and written Task 1	0.49	0.57	0.46	2.76*
Hands-on Task 2 and written Task 2	0.49	0.57	0.52	0.66
Hands-on multiple choice and written multiple choice	0.49	0.57	0.46	2.76*
Hands-on justification and written justification	0.59	0.71	0.6	1.78
Hands-on and non-analogous parts of written test				
Hands-on total and written identification	0.5	0.47	0.56	1.11
Hands-on total and written prediction	0.51	0.48	0.59	1.38

* = significant at .05 level.

the test, and correlation between scores on the first task, but not on the second task. Thus, taking the hands-on test first rendered total scores and some subscores more consistent across hands-on and written analogues. However, the positive impact of the hands-on experience on the stability of scores was confined to tasks that had highly similar structures; correlations between hands-on scores and scores on the non-analogue parts of the written test (identification and prediction) were not higher if the hands-on test was taken first. In fact, the correlations were higher if the written test was taken first, but the differences between correlations were not large enough to be statistically significant.

Order effects on reliability/generalizability. Table 6 shows the estimated univariate generalizability (G) coefficients ($\hat{\rho}$), phi coefficients ($\hat{\Phi}$), and variance components for scores on hands-on and written analogue tests when the test was taken on the first day and when it was taken on the second day of the study. The G coefficient shows the dependability (reliability) of the relative ordering of examinees (a norm-referenced interpretation of test scores); the phi coefficient shows the dependability of the absolute level of an examinee's performance independent of others' performance (cf. criterion-referenced interpretations). The G-study design used was p x (c x i):t, where p stands for person, c stands for concept being measured (voltage, resistance, or current), i stands for item type (multiple-choice or justification), and t stands for task. Each person got two tasks;

Table 6

Estimated Variance Components and Generalizability Coefficients From Univariate Generalizability Analyses for Hands-On and Written Analogue Tests, p x (c x i):t Design

Test	Facet	First day	Second day
Hands-on	p	0.03867	0.05203
	t	0	0
	c(t)	0.00302	0.0034
	i(t)	0.01113	0.0098
	pt	0.00004	0
	pc(t)	0.06127	0.04982
	pi(t)	0.01616	0.00313
	ci(t)	0.00648	0.01326
	pci(t),e	0.0842	0.08872
	$\hat{\Phi}$	0.61	0.72
	$\hat{\rho}$	0.64	0.76
Written analogue	p	0.03554	0.06215
	t	0	0
	c(t)	0.00693	0.00534
	i(t)	0.00833	0.00361
	pt	0	0
	pc(t)	0.05542	0.04935
	pi(t)	0.01896	0.02204
	ci(t)	0.00301	0.00748
	pci(t),e	0.08997	0.08823
	$\hat{\Phi}$	0.59	0.73
	$\hat{\rho}$	0.62	0.75

therefore tasks are crossed with persons. Nested within each task are three concepts (voltage, resistance, and current), which are crossed with two item types (multiple-choice and justification), because there is one multiple-choice and one justification question for each concept. Concepts and item types were nested in tasks, because the different set of equipment for the two tasks made the items different. All facets were treated as random because the concepts measured as well as the item types and tasks were considered to be drawn from the universe of all possible tasks, item types, and concepts in the domain of electric circuits. Generalizability and phi coefficients for tests taken on the first day (ranging from

.59 to .64) were lower than those for tests taken on the second day (ranging from .72 to .76). Thus, generalizability was higher on whichever test was taken second.

Table 7 shows the univariate G coefficients and variance components for scores on the written identification and prediction items. For these analyses, the design used was simply p x i, that is, persons crossed with items. Generalizability for these types of items was higher when the written test was taken on the second day, that is, after the hands-on test (.62 and .72 on Day 1, versus .71 and .85 on Day 2). Thus, overall, generalizability was greater for whichever items were taken on the second day of testing. It appears that the practice on the first day gave students a more stable knowledge base from which to operate on the second day.

The higher generalizability on the second day was due to higher universe-score variance rather than lower error variance (see Table 6). Inspection of the data revealed a wider distribution of scores on the test taken on the second day than on the test taken on the first day. Means and standard deviations were

Table 7

Estimated Variance Components and Generalizability Coefficients From Univariate Generalizability Analyses for Written Identification and Prediction Items, p x i Design

Items	Facet	First day	Second day
Identification items		(n = 177)	(n = 156)
	Persons (p)	0.052	0.069
	Items (i)	0.007	0.013
	p x i,e	0.189	0.169
	$\hat{\Phi}$	0.62	0.69
	$\hat{\rho}$	0.62	0.71
Prediction items		(n = 169)	(n = 155)
	Persons (p)	0.024	0.054
	Items (i)	0.038	0.026
	p x i,e	0.169	0.172
	$\hat{\Phi}$	0.67	0.83
	$\hat{\rho}$	0.72	0.85

higher on the second day, and there were fewer very low scores and more very high scores on the second day. For example, 9% of students who took the hands-on test on Day 1 got a mean score lower than .10, but only 4% of students who took the hands-on test on the second day got a mean score lower than .10. Only 6% of students who took the hands-on test on Day 1 got a score higher than .80, but 20% of students who took the hands-on test on Day 2 got a score higher than .80.

An examination of the Day 2 scores of students who scored in each decile on their Day 1 test revealed that it was the low-scoring students on Day 1 who improved their scores the most on Day 2, regardless of which test was taken first. Table 8 shows the means and standard deviations for Day 1 and Day 2 scores of students who scored in each decile on Day 1. Students who scored in the top three deciles on Day 1 did not improve their scores on Day 2. However, students who scored in the bottom three deciles on Day 1 showed a big improvement on Day 2. For example, students who scored lower than .1 on the hands-on test on Day 1 had a mean score of .21 on the written analogue test on Day 2. Students who scored between .1 and .19 on the written analogue test on Day 1 had a mean score of .47 on the hands-on test on Day 2.

Table 8

Change in Scores From Day 1 to Day 2 for Day 1 Deciles

	Hands-on test on Day 1 (n = 158)					Written test on Day 1 (n = 186)				
		Day 1 scores (hands-on)		Day 2 scores (written analogue)			Day 1 scores (written analogue)		Day 2 scores (hands-on)	
Interval	n	Mean	*SD*	Mean	*SD*	n	Mean	*SD*	Mean	*SD*
.9-1	3	.93	.02	.91	.02	16	.93	.02	.90	.04
.8-.89	7	.85	.04	.82	.10	14	.97	.03	.76	.14
.7-.79	15	.75	.03	.73	.15	31	.76	.03	.76	.14
.6-.69	22	.66	.03	.75	.17	31	.66	.02	.77	.20
.5-.59	19	.55	.03	.60	.22	28	.54	.03	.69	.15
.4-.49	22	.45	.03	.50	.23	24	.45	.03	.71	.24
.3-.39	17	.36	.02	.40	.21	14	.35	.03	.52	.27
.2-.29	18	.25	.03	.42	.22	12	.26	.03	.45	.23
.1-.19	21	.16	.03	.26	.18	12	.17	.02	.42	.21
0-.9	14	.04	.03	.21	.19	4	.04	.04	.47	.14

Summary of order effects. The order in which students took the written and hands-on tests had a significant impact on mean performance, on univariate generalizability, and on correlations within and across tests. Specifically, the following order effects were found:

- performance was higher on hands-on and written analogues, but not on identification and prediction items, on the second day;
- the greatest improvement in scores from Day 1 to Day 2 occurred for students who had low scores on Day 1;
- scores were more consistent (generalizable) on whichever test was taken on the second day;
- correlations between scores on written and hands-on tests were higher if the hands-on test was taken first.

Because of these order effects, all further analyses were conducted for each order separately.

Interchangeability of Multiple-Choice and Justification Item Types

In Table 3, we reported that the correlations between the mean of all multiple-choice items (averaging over questions and tasks) and the mean of all justification items were quite high for the hands-on test (r = .73) and the written analogue (r = .62). These results suggest that reporting separate scores for multiple-choice and justification items may not be justifiable. We will now examine the interchangeability of multiple-choice and justification scores in more detail. First we compare the observed correlations between multiple-choice and justification items with the observed correlations between comparable items that vary by task, test format, and concept. Second, we examine disattenuated correlations between multiple-choice and justification items.

Observed correlations between items. Analyses were first conducted at the item level to determine the magnitude of the correlations between multiple-choice and justification scores for the same question on the test (e.g., "Which circuit had higher voltage?" and "Why?") and whether these correlations were higher than those among items that came from different tasks, different test formats, or measured different concepts. Table 9 summarizes the correlations between items that varied only in item type (multiple-choice, justification), only by task (Task 1, Task 2), only by test format (hands-on, written), or only by concept (voltage, resistance, current).

Table 9

Range in Correlations Between Items Varying in Item Format, Task, Test Format, or Concept

Source of variation between items	Hands-on test on Day 1 (n =158)			Written test on Day 1 (n = 186)		
	Voltage	Resistance	Current	Voltage	Resistance	Current
Item type (multiple-choice, justification)[a]	.52 to .74	.51 to .64	.13 to .28	.53 to .79	.51 to .64	.13 to .40
Task (Task 1, Task 2)[b]	.31 to 54	.35 to .54	.16 to .64	.31 to .52	.26 to .40	.16 to .39
Test format (hands-on, written)[c]	.31 to .41	.38 to .48	.14 to .39	.17 to .48	.21 to .39	.16 to .39
Concept (voltage, resistance, current)[d]	.17 to .40	.17 to .42	.19 to .40	.09 to .46	.11 to .46	.14 to .51

[a] Items come from the same test, the same task, and the same concept; the only difference is the item type.

[b] Items come from the same test, the same concept, and use the same item type; the only difference is the task.

[c] Items come from the same task, the same concept, and use the same item type; the only difference is the test format.

[d] Items come from the same test, the same task, and use the same item type; the only difference is the concept.

As can be seen in Table 9, for voltage and resistance, the correlations between the multiple-choice and justification scores (first row of the table) were substantial and were higher than correlations among items from different tasks (second row), test format (third row), or concept (fourth row). For example, among students who took the hands-on test on the first day, the correlations between multiple-choice and justification scores for the voltage question ranged from .52 to .74 across the different tasks and tests. In contrast, the correlations between an item pertaining to voltage on Task 1 of a test (e.g., "Which circuit had higher voltage?") and the same item pertaining to voltage on Task 2 on the same test and using the same item type were lower, ranging from .31 to .54. Moreover, the correlations between a voltage item on Task 1 of the hands-on test and the same item on Task 1 of the written analogue test and using the same item format were also lower, ranging from .31 to .41. Finally, the correlations between a voltage item (e.g., "Which circuit had higher voltage?") and items that measured other tasks but came from the same task, the same test, and used the same item

format (e.g., "Which circuit had higher resistance?") were also lower, ranging from .17 to .40.

The same pattern of correlations held for items measuring resistance but not for those measuring current. For items measuring current, the correlations between multiple-choice and justification items were considerably lower than those for voltage and resistance (range for current was .13 to .40, whereas for voltage the range was .52 to .79). This is because the concept of current is inherently more complex than voltage and resistance, and the scoring of current justifications depended on reference to both the voltage and resistance in the circuits being compared (see earlier section on scoring). Consequently, the justification questions for current were more difficult relative to the multiple-choice questions than were the justification questions for voltage and resistance. This hypothesis is confirmed by the mean scores for multiple-choice and justification items for each concept. Combined for both written and hands-on tests, the mean multiple-choice scores for voltage, resistance, and current were .56, .55, and .55 respectively, whereas the mean justification scores were .48, .48, and .26. Often, students made reference to only one aspect of the circuits that determined their relative current, that is, a reference to either the voltage or the resistance, which automatically lowered the rating of the response.

Correlations between scores for items measuring the concept of current across the two tasks or on the same item type across the hands-on and written versions of the tasks were not generally lower than those correlations for voltage and resistance. Also, correlations between scores on the same item type across concepts were similar for all three concepts. Thus, the discrepancy in correlations for current, compared with those for voltage and resistance, was confined to correlations between multiple-choice and justification item types. This suggests that for concepts that are more complex, multiple-choice items may not reveal misconceptions that open-ended responses uncover.

The patterns of correlations (at least for voltage and resistance) support the contention that students' scores on multiple-choice and justification items were closely related and that the relationship was stronger than correlations between items that differed by task, test format, or concept. It should be noted that these results relate to norm-referenced interpretations of scores, not criterion-referenced interpretations. We cannot assert that the absolute performance of students on each item type was the same. Although we cannot equate the

multiple-choice and justification scales (the multiple-choice responses were scored dichotomously and the justifications were rated on a 4-point scale), the mean performance on multiple-choice items was higher (for example, .58 for the hands-on test) than performance on justification items (.41 for the hands-on test).

Further evidence that absolute performance on the multiple-choice items was higher than on the justification items can be found by examining the justification scores of students who got a multiple-choice item correct. Taking the voltage item as an example, 21% of students who answered the multiple-choice item correct got a score of 0 for their written justifications, and only 60% of students with correct multiple-choice answers got a perfect score of 1 (on the 4-point scale) on the justification item. In contrast, 75% of students who got the multiple-choice item wrong scored 0 on the justification item, but 12% of them got a perfect score of 1 on their justifications.

While the observed correlations suggest that the multiple-choice and justification item types are strongly related, the observed correlations are partly dependent on reliability of the multiple-choice and justification measures. Consequently, we disattenuated the correlations to determine whether the correlation between true scores on multiple-choice and justification is close to unity. The correlations that we disattenuated were those between mean scores across all items of the same item type within a test (e.g., the correlation between the mean of all multiple-choice scores on the hands-on test and the mean of all justification scores on the same hands-on test).

Disattenuated correlations. To calculate disattenuated correlations, we carried out multivariate generalizability analyses. Because of the relatively infrequent use of this technique we will explain in detail here the statistical theory and procedure for this analysis. 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 multiple-choice scores on the hands-on test. There are three multiple-choice questions for the first task and three multiple-choice questions for the second task. Thus, the design is $p \times q{:}t$ or, in words, persons

crossed with questions nested within tasks. The total variance of the observed scores equals the following sum of variance components:

$$\sigma^2(X_{pq:t}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(q:t) + \sigma^2(pt) + \sigma^2(pq:t,e) . \quad (1)$$

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 *m* for multiple-choice and *j* for justification, the total variance of the multiple-choice scores and the total variance of the justification scores can be decomposed as follows:

$$\sigma^2({}_mX_{pq:t}) = \sigma^2(mp) + \sigma^2(mt) + \sigma^2(mq:t) + \sigma^2(mpt) + \sigma^2(mpq:t,e) .$$

$$\sigma^2({}_jX_{pq:t}) = \sigma^2(jp) + \sigma^2(jt) + \sigma^2(jq:t) + \sigma^2(jpt) + \sigma^2(jpq:t,e) . \quad (2)$$

Multivariate generalizability theory (Brennan, 1992; Brennan, Gao, & Colton, 1995; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1981; Webb, Shavelson, & Maddahian, 1983) decomposes covariances among scores as well as variances among scores. In the present study, the total covariance between multiple-choice scores and justification scores is decomposed into the following components of covariance:

$$\sigma({}_mX_{pq:t},{}_jX_{pq:t}) = \sigma(mp,jp) + \sigma(mt,jt) + \sigma(mq:t,jq:t)$$

$$+ \sigma(mpt,jpt) + \sigma(mpq:t,e,jpq:t,e) . \quad (3)$$

The covariance component $\sigma(mp,jp)$ is the covariance between persons' universe scores for multiple-choice and justification. The remaining terms in Equation 3 are error covariance components. The disattenuated correlation (see Brennan et al., 1995; Cronbach et al., 1972, p. 287) between multiple-choice and justification scores is

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

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.

Using these procedures, the estimated universe-score variance and covariance components for students who were administered the hands-on test on Day 1 are:

$$\hat{\sigma}(mp,jp) = .0341$$

$$\hat{\sigma}^2(mp) = .0462$$

$$\hat{\sigma}^2(jp) = .0401$$

The substantial estimated covariance component relative to the estimated variance components shows that students with high universe scores on multiple-choice tended to have high universe scores on justification. The disattenuated correlation using Equation 4 is .79. This is the estimated value of the correlation between multiple-choice and justification scores as the number of tasks and the number of items per task approach infinity. The disattenuated correlations between multiple-choice and justification scores on the hands-on and written analogue tests on Day 1 and Day 2 range from .73 to .94 (see Table 10), compared to a range of .60 to .77 for observed correlations. Although one of the disattenuated correlations is close to unity, the remainder are not. These results suggest that scores on the two item types are generally not correlated strongly enough to be considered interchangeable. However, given the difference in costs of scoring multiple-choice and open-ended justification responses, some policy makers might consider this level of consistency high enough to justify using tasks where only responses to multiple-choice items are recorded and scored in large-scale assessment.

Table 10

Observed and Disattenuated Correlations: Multiple-Choice and Justification Items

	Hands-on		Written analogue	
Test taken first	*r*	*r**	*r*	*r**
Hands-on first (*n* = 158)	0.65	0.79	0.62	0.73
Written first (*n* = 186)	0.77	0.94	0.60	0.75

Note. *r* = observed correlation; *r** = multivariate G disattenuated correlation.

Given that the two item types are strongly related but not interchangeable, we decided not to drop an item type and instead calculated the mean of the multiple-choice and justification scores for each question and used the mean score for each concept in all remaining analyses.

Interchangeability of Hands-On and Written Analogue Tests

The extent to which performance on science assessments is affected by having the opportunity to manipulate equipment was explored by examining mean performance and consistency of performance across hands-on and written analogues. The results presented in Table 1 indicated that mean scores on the hands-on and written analogue tests were very similar. Thus, there was no effect of equipment manipulation on average performance.

The observed correlations between hands-on and written analogue scores (shown in Table 5) showed a strong relationship for students who took the hands-on test first (.71) but only a moderate relationship for students who took the written test first (.58). Disattenuated correlations using the multivariate generalizability procedures described above are presented in Table 11. The

Table 11

Observed and Disattenuated Correlations: Hands-On Tasks and Written Analogues

Test taken first	*r*	*r**
Hands-on first (*n* = 158)	0.71	0.96
Written first (*n* = 186)	0.58	0.76

Note. *r* = observed correlation; *r** = multivariate G disattenuated correlation.

disattenuated correlation between hands-on and written analogue scores was close to unity among students who took the hands-on test first (.96), but was not close to unity among students who took the written test first (.76). These results suggest that, as the number of tasks and the number of items within tasks approach infinity, the two test formats may be interchangeable if students are administered the hands-on test first but not if students are administered the written test first. This raises doubts about being able to use the written analogue as an acceptable substitute for the hands-on test. If only the written test is administered, the scores would *not* be very similar to those that students would obtain if they had the opportunity to manipulate equipment.

It should be noted that the correlations reported here between hands-on and written analogue scores (both observed and disattenuated correlations) are generally higher than the correlations between hands-on and non-hands-on methods found in previous research. The high correlations found in this study can be attributed to fact that the hands-on and non-hands-on tests had analogous structures—indeed, all aspects of the two tests were identical except for the presence of the equipment—and were designed to tap similar knowledge and cognitive processes. Without such tightly matched features of the tests, we would expect the correlations to be lower.

Interchangeability of Written Non-Analogues With the Hands-On Test

If one considers hands-on performance to be the "benchmark" performance in science, then we can ask "To what extent can other kinds of assessments act as surrogates and yield similar estimates of performance?" The results presented in the previous section indicate that performance on a written analogue of a hands-on task was a relatively strong predictor of hands-on performance although the two tests were only interchangeable if the benchmark was administered before the surrogate. We will now examine the relationship between hands-on performance and performance on the identification and prediction items on the written test. Tables 12 and 13 present the observed and disattenuated correlations between hands-on scores and identification and prediction scores. The observed correlations ranged from .41 to .59 and the disattenuated correlations range from .63 to .81. These correlations show that, although written non-analogue scores were fairly strong predictors of hands-on scores, the relationships were not strong enough for the written non-analogue tests to be considered a suitable surrogate for the hands-on test.

Table 12

Observed and Disattenuated Correlations: Hands-On Test and Identification Items

Test taken first	*r*	*r**
Hands-on first (*n* = 158)	0.47	0.67
Written first (*n* = 186)	0.56	0.81

Note. *r* = observed correlation; *r** = multivariate G disattenuated correlation.

Table 13

Observed and Disattenuated Correlations: Hands-On Test and Prediction Items

Test taken first	*r*	*r**
Hands-on first (*n* = 158)	0.48	0.63
Written first (*n* = 186)	0.59	0.81

Note. *r* = observed correlation; *r** = multivariate G disattenuated correlation.

Summary of Interchangeability Results

Taking the hands-on test as the benchmark, a written analogue version of the test was found to generate highly similar scores if the written analogue was taken AFTER the hands-on test. If the written analogue was taken without first having the benefit of hands-on manipulation of equipment, then the written analogue test did not generate scores that were sufficiently similar to hands-on scores to render the tests interchangeable. Other types of written items were found to be not interchangeable with the hands-on test. Thus, we are forced to conclude that even the most closely matched written analogue is not interchangeable with a test that permits hands-on manipulation of equipment. Within either a hands-on or written analogue test, scores on multiple-choice and justification items related to the task are highly correlated, and although the items are not exactly interchangeable, it could be argued that the disattenuated correlations are high enough (range: .73 to .94) to warrant elimination of costly open-ended items in large-scale testing.

Discussion

This study set out to investigate the interchangeability of different types of assessment methods for measuring middle-school students' understanding of science concepts**.** The one previous study that had the most similar goal (Baxter & Shavelson, 1994) found that only scores directly based on aspects of hands-on performance (either observation scores or scores of notebook entries students kept during hands-on work) were interchangeable. Scores on other forms of assessment such as short-answer questions or multiple-choice questions yielded different patterns of scores than did scores based on hands-on performance. The authors concluded that "each method may measure different yet related aspects of science achievement" (p. 297). The study reported here used a more complicated design to further probe issues related to assessment methods' interchangeability. This study differed from Baxter and Shavelson's study in the following ways:

1. the domain of science sampled was narrower;
2. a written test that was completely analogous to the hands-on test, but without access to equipment, was constructed so that the unique contribution of hands-on manipulation could be examined;
3. both the hands-on and written tests were administered over a 2-day period, whereas in Baxter and Shavelson's study the different assessments were administered over the course of an entire semester at 3- to 4-week intervals;
4. actual hands-on performance was not scored; instead diagrams drawn after hands-on work and responses to related multiple-choice and justification items were scored;
5. the order of administration of hands-on and written tests was counterbalanced; this permitted examination of the effects of order of assessment methods on performance. In Baxter and Shavelson's study, all students took the assessments in the same order, with the hands-on test coming after the written tests (short-answer and multiple-choice).

In spite of the design differences, this study confirmed some of the conclusions of Baxter and Shavelson's study. First, we found that scores on two types of items related to hands-on tasks were highly correlated. In our case, the high correlations were between multiple-choice and justification items about the product of the hands-on activity; in Baxter and Shavelson's case, the high

correlations were between observation scores and scores of written work related to the hands-on task. A general hypothesis that might be posed from these results is that almost any aspect of hands-on performance that can be scored will yield similar norm-referenced results for an individual student. If that is the case, then large-scale assessment could justify sacrificing the scoring of actual performance or open-ended responses related to hands-on tasks, and instead score responses to multiple-choice questions that are based on the hands-on activity.

In this study, absolute performance on multiple-choice and justification items was not comparable. Scores on multiple-choice items were generally higher than scores on justification items. However, some students who scored 0 on a multiple-choice item obtained a perfect score on the corresponding justification item. Similarly, some students who got a multiple-choice item correct scored 0 on the corresponding justification item.

One reason for the higher than usual correlations between multiple-choice and justification questions found in this study is the fact that the questions were linked to very specific concepts. If the domain being sampled were broader and we compared multiple-choice questions measuring some topic areas with justification questions measuring other topic areas, we would not find such high correlations. Therefore, we are not recommending that open-ended item formats be jettisoned altogether; only in cases where there is redundancy in the knowledge being measured by the two formats.

The second finding of the Baxter and Shavelson study that was confirmed by this study is the importance of hands-on manipulation in eliciting stable estimates of knowledge. Baxter and Shavelson based their advocacy of hands-on tasks on the fact that their hands-on tasks generated scores that were *not* highly correlated with scores from written tests. They concluded that hands-on tasks were therefore tapping some unique aspect of science knowledge; the hands-on scores also had the highest reliabilities (internal consistency represented by relative generalizability coefficients) of the assessment methods used in the study. In this study, we found that when the hands-on test was taken before the written test, scores on the written analogue test were more highly correlated with scores on the hands-on version than if the written analogue was taken before the hands-on test. It appears that the opportunity to manipulate and get dynamic

feedback from real equipment leads to more stable estimates of student knowledge.

The fact that correlations between scores on hands-on and written analogue tasks with identical structure and substance were only moderate when the written test was taken first (observed r = .58; disattenuated r = .76) indicates that eliminating the hands-on component inherently changes the cognitive processes induced by completing the task. Think-aloud studies are needed to explore the exact nature of these differences. This study also found that scores on hands-on tasks were not highly correlated with scores on non-analogous multiple-choice items that were designed to measure students' ability to (a) identify the relative voltage, resistance, and current in different circuits and (b) predict the effects of circuit changes on voltage, resistance, and current. These correlations were also higher if the hands-on test was taken first, although not sufficiently higher for the difference to be statistically significant.

Regardless of which test was taken first, performance on the version of the test (hands-on or written analogue) that was taken on the second day was higher. Scores were also more reliable (generalizable) on the second day. Students who had lower scores on the first day showed the most improvement on the second day. This practice effect was evident regardless of which test was taken first. Thus, less able students could benefit greatly from the opportunity to practice tasks and items similar to those that will be on a real test.

The correlations between the two types of multiple-choice items that were not associated with hands-on or written analogue tasks, that is, the prediction and identification items on the written test, were relatively low, and these correlations were affected by order of administration. When the written test was taken second, the correlation between the identification and prediction items was .68, but it was only .51 when the written test was taken first. It appears that hands-on experience may stabilize knowledge across the board by giving students an opportunity to refine and tune their knowledge, which they can then call on in a variety of testing situations.

The large order effects found in this study have implications for the design of future investigations of interchangeability of assessment methods. The order in which students take a battery of assessments may influence the patterns of correlations that will be found. If all students in this study had taken the hands-

on test first, we would probably have concluded that hands-on and written analogue tests are interchangeable. If all students had taken the written test first, we would have concluded the opposite. Future studies should either counterbalance order of administration, or else should administer the "benchmark" assessment after the surrogates because it is the correlation when a surrogate is taken first that matters. If the benchmark is administered first, it may influence students' performance on the surrogate. Fortunately, in the Baxter and Shavelson study, the benchmark hands-on assessment was administered after the written surrogates.

The results of this study reinforce the conclusion that it is difficult to create items and tasks that have similar cognitive demands and lead to similar estimates of individual student knowledge. Our results show that hands-on and written tests that have analogous content and structure, but that lack the hands-on component, are not interchangeable. However, certain item types within the context of hands-on or written extended tasks may be interchangeable, for example, multiple-choice and justification items, depending on the criterion one uses for interchangeability.

References

Baxter, G. P., Glaser, R., & Raghavan, K. (1993). *Cognitive analysis of a science performance assessment* (CSE Tech. Rep. No. 382). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baxter, G. P., & Shavelson, R. J. (1994). Performance assessments: Benchmarks and surrogates. *International Journal of Educational Research, 21,* 279-298.

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.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*, 253-271.

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

Dunbar, S. B., Koretz, D., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, 289-304.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.

Hamilton, L. S., Nussbaum, M., & Snow, R. E. (1995, April). *Alternative interview procedures for validating science assessments.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33*, 71-92.

Nichols, P., & Sugrue, B. (1997). *Construct-centered test development for NAEP's short forms.* Paper commissioned by the National Center for Educational Statistics. Iowa City: University of Iowa, College of Education.

Royer, J. M, Cisero, C. A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*, 201-243.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207-217.

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. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133-166.

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice, 16*(3), 16-25.

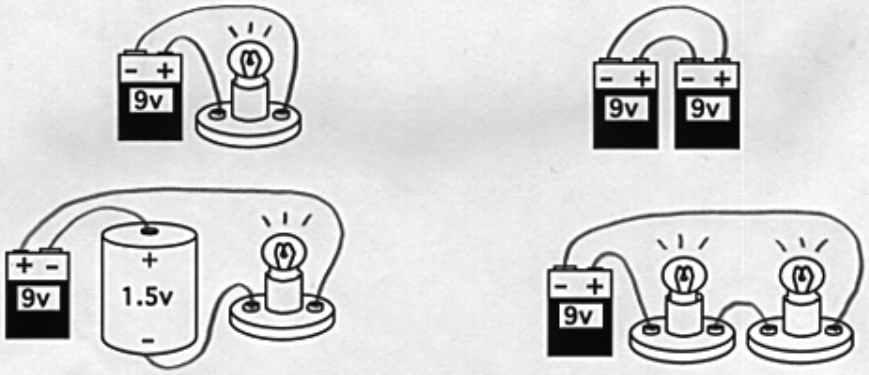
Webb, N. M., Shavelson, R. J., & Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Fyans (Ed.), *Generalizability theory: Inferences and practical applications* (pp. 67-81). San Francisco, CA: Jossey-Bass.

Appendix

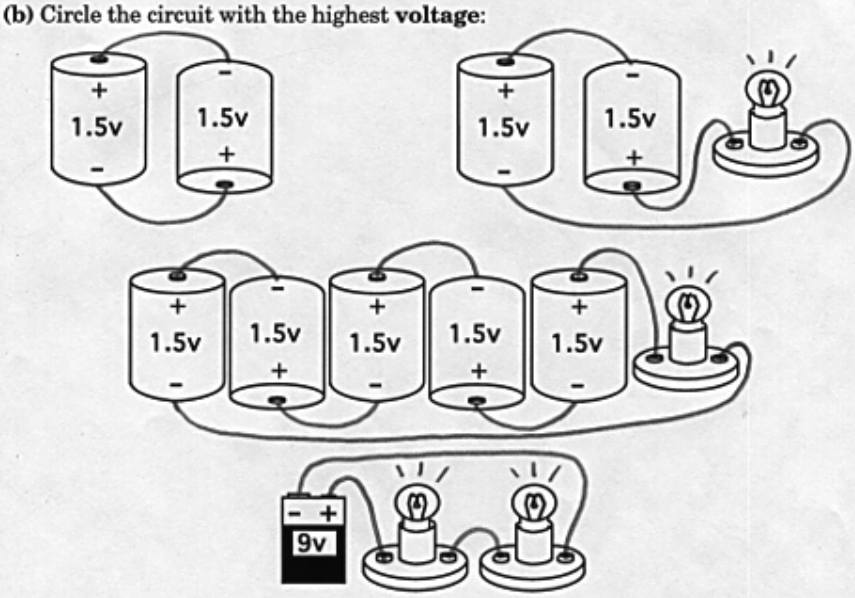
Copies of Test Items

2. For each of the following two sets of circuits, (a) and (b), circle the circuit that has the **highest voltage**. Assume that all circuits are properly connected.

(a) Circle the circuit with the highest **voltage**:

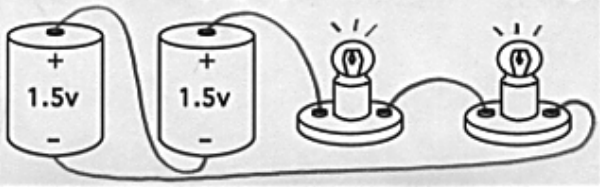


(b) Circle the circuit with the highest **voltage**:



---

6. Predict what will happen to the voltage, resistance, and current in the following circuit if each of the changes listed in the chart is made. Circle **INCREASE, DECREASE,** or **NO CHANGE,** in each box in the chart. Assume that the circuit is properly reconnected after a change is made.

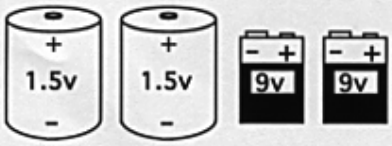


What will happen if you	Voltage	Resistance	Current
add another bulb?	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE
add a 9-volt battery?	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE
remove one bulb?	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE
remove one battery?	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE
add a glass rod?	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE
remove both bulbs?	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE	INCREASE DECREASE NO CHANGE

13. (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)

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

_____

_____

_____

13

13. (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.)

_____

_____

_____

13. (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.)

_____

_____

_____

13. (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.)

_____

_____

_____