

**On the “Exchangeability” of Hands-On and  
Computer-Simulated Science Performance Assessments**

CSE Technical Report 531

Anders Rosenquist, Richard J. Shavelson,  
and Maria Araceli Ruiz-Primo  
CRESST/Stanford University

November 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  
Richard J. Shavelson, Project Director, CRESST/Stanford University

Copyright © 2000 The Regents of the University of California

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 Office of Educational Research and Improvement, or the U.S. Department of Education.

# ON THE “EXCHANGEABILITY” OF HANDS-ON AND COMPUTER-SIMULATED SCIENCE PERFORMANCE ASSESSMENTS

**Anders Rosenquist, Richard J. Shavelson, and Maria Araceli Ruiz-Primo**  
**CRESST/Stanford University**

## **Abstract**

Inconsistencies in scores from computer-simulated and “hands-on” science performance assessments have led to questions about the exchangeability of these two methods (e.g., Baxter & Shavelson, 1994), in spite of the former’s highly touted potential (e.g., Bennett, 1999). Five possible explanations of students’ inconsistent performances were considered: (1) inadequate exposure to computers and simulations, (2) differential views of computer-simulated (2-dimensional icons) and hands-on tasks, (3) different methods tapping different aspects of achievement, (4) partial or incomplete knowledge, and (5) a combination of partial knowledge and method differences. The first explanation was ruled out by the fact that students had computers in their classes and used them for a variety of purposes, including simulation. The second explanation was ruled out using talk-aloud data, randomized experiments, and student questionnaire responses. If explanation 3 were tenable, the correlation between Electric Mysteries scores at time 1 and time 2 for either hands-on or computer simulation should be higher than the correlation between hands-on scores and computer simulation scores at either point in time. Shavelson, Ruiz-Primo, and Wiley (1999) provided correlations that did not jibe with this expectation. To explore the remaining two possible explanations dealing with student expertise, we compared the performance of high school physics students (“experts”) to that of Baxter and Shavelson’s elementary school students and found, somewhat surprisingly, that these “experts” were far from expert. Indeed, they were no more expert than the elementary students. Consequently, we have narrowed the possible explanations for the lack of exchangeability between computer-simulated and hands-on performance assessments to one of two choices: partial knowledge or the interaction of partial knowledge with method. The jury is still out.

Two critical factors have contributed to the slow adoption of science performance measures in large-scale assessments: (a) high per pupil costs (Stecher & Klein, 1997), and (b) development and administration time (Solano-Flores & Shavelson, 1997). One possible solution to the cost and time problems may be computer-based versions of hands-on assessments (Pine, Baxter, & Shavelson, 1993). Yet, there are certain trade-offs when hands-on tasks are translated into an electronic environment that might compromise quality. Are the hands-on and computer

versions the same, even when they “look alike?” Are scores earned in one medium the same as—“exchangeable for”—scores earned in the other?

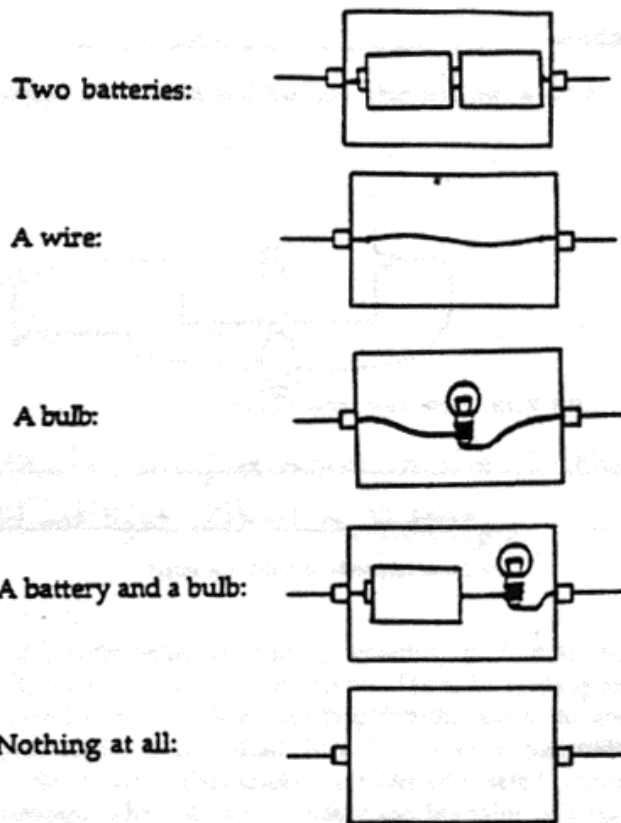
To address these questions, Shavelson, Baxter, and Pine (Baxter, 1995; Baxter & Shavelson, 1994; Pine et al., 1993; Shavelson, Baxter, & Pine, 1991) compared the same 5th- and 6th-grade students’ scores earned on two hands-on and computer-based performance assessments, “Bugs” and “Electric Mysteries.” For both assessments, they reported moderate correlations ( $\sim .53$ ) between the two methods ( $N > 300$ ). The correlations were sufficiently low to lead Baxter and Shavelson (1994) to question the exchangeability of the two methods.

More specifically, Shavelson, Baxter, Pine, and Yure (1991) reported a correlation of .84 between *direct observation*, where a rater scored a student’s performance of a hands-on investigation in real time, and *notebook*, where a rater scored a student’s notebook based on the *same* investigation. One month following the hands-on investigation, the same students performed the computer simulation. Here, the correlations between scores based on either the observation and the computer simulation or the notebook and the computer simulation were about .53, compared to the observation/notebook correlation of .84. By changing methods and separating the methods by a month, the correlation dropped from .84 to .53. Because the computer simulations were developed to be identical to their hands-on counterparts, one might reasonably expect or predict a greater correlation between the two methods. Shavelson, Baxter, and Pine (1991) concluded that hands-on and simulated investigations were not tapping the same knowledge.

The purpose of the current study was to address the issue of exchangeability of hands-on and computer-simulated performance assessments, in particular the Electric Mysteries assessment. Electric Mysteries is an electrical circuits investigation in which students determine the contents of six “mystery boxes” by connecting wires, batteries, and light bulbs to the boxes (Figure 1). The boxes may contain a wire, a bulb, two batteries, a battery and bulb, or nothing. Students are not allowed to open the boxes. The investigation can be performed using either actual boxes, wires, and bulbs that can be physically manipulated, or a computer simulation with box, wire, and bulb representations that can be manipulated on-screen using a mouse.

Several factors may differentiate the hands-on and computer simulation versions of this assessment.

Find out what is in the six mystery boxes A, B, C, D, E and F. They have five different things inside, shown below. Two of the boxes will have the same thing. All of the others will have something different inside.



For each box, connect it in a circuit to help you figure out what is inside. You can use your bulbs, batteries and wires any way you like.

Figure 1. Electric Mysteries investigation.

1. One possible explanation for the moderate correlation is that students taking the hands-on and computer-simulated assessments may not have been adequately instructed in the use of computer simulations since their science curricular activities were primarily hands-on.<sup>1</sup> Therefore, students may not have been completely comfortable in the computer simulation environment, which in turn may have affected their performance.

2. A second possible explanation may be that, although the computer version was built to “look and feel” the same as a hands-on assessment, students might have viewed each method differently.

<sup>1</sup> The students all had access to a Macintosh computer and were familiar with its operation and use.

3. A third, related hypothesis is that each method might afford the user a unique and different experience that may tap different aspects of achievement leading to inconsistencies in performance across methods.

4. A fourth possible explanation for only moderate correlations between hands-on and computer-simulated performance assessment scores is that issues of exchangeability are not due to method, but to students' partial knowledge within the specific science domain. Their performance is inconsistent not from one method to another, but from one occasion to another. Each time they encounter the investigation, they approach it slightly differently, much as novices do (as in the expert-novice literature, e.g., Mestre, 1994). That is, because students' procedural and content knowledge is limited, the features of one or another method catch a student's eye and performance varies, even though more expert students would recognize that each method or occasion is based on the same underlying scientific principles and thus perform more similarly on repeated versions of the investigation.

5. A fifth possible explanation combines both the medium and expertise explanations. Because methods vary, students' performance varies. Because students are not experts, their performance varies from one occasion to another. When both occasion and method vary, performance of students with partial knowledge will vary considerably.

To empirically test explanations 2-5, we would need data from a study in which hands-on and computer-simulated versions of the Electric Mysteries assessment were both administered to the same students on two different occasions (see Figure 2). If the medium explanations (2 and 3) were correct, the correlations between the two methods, ( $r_{H1C1}$  or  $r_{H2C2}$ ), would be substantially lower than the correlations between the same methods (hands-on/hands-on and computer/computer) lagged over time ( $r_{H1H2}$  and  $r_{C1C2}$ ). If the partial-knowledge explanation (4) were correct, all correlations, regardless of method or occasion would be of the same magnitude (about .53).

If the medium/expertise explanation (5) were correct, the correlation across methods at time 1 ( $r_{H1C1}$ ) or at time 2 ( $r_{H2C2}$ ) should be higher than the correlation between different methods lagged over time (i.e.,  $r_{H1C2}$ ,  $r_{H2C1}$ ). Likewise, the correlation between the same measure at two time points ( $r_{H1H2}$ ,  $r_{C1C2}$ ) should be higher than the cross-lagged correlations ( $r_{H1C2}$ ,  $r_{C1H2}$ ). Unfortunately, we do not have

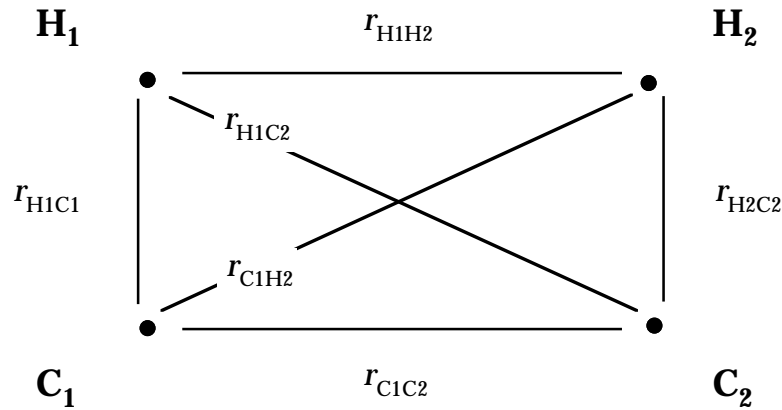


Figure 2. Possible correlations between hands-on and computer simulation performance scores taken.

data for the computer simulation at two points in time and so cannot directly test hypothesis 5.

We begin our search for a solution to the exchangeability problem by addressing explanations 1 and 2: Students were not adequately instructed in the use of the two methods, and although the methods looked alike to designers, students did not view them that way. To investigate the moderate correlation between hands-on and computer simulation methods reported by Shavelson, Baxter, and Pine (1991), Ruiz-Primo (unpublished work) conducted several short studies with the Electric Mysteries assessment in 5th-grade classrooms in Montecito, California. First, the instructions to the computer-simulated investigation used by Shavelson, Baxter, and Pine (1991) were changed because some of the students took a long time to read them. The intent in changing the instructions was to make them an easier and less time consuming to read, not to change performance by instruction. After several iterations, easily interpretable instructions were created.

In the next study, students' ability to interpret the various computer-generated icons was investigated. Two short surveys asked students to identify various icons and the relative intensity of the light bulbs. Very few students had problems with icon interpretations. Of 17 students, 15 interpreted all icons perfectly.

The third study concerned the exchangeability of the hands-on and computer simulation methods. In a randomized experiment, half of the students used the hands-on boxes as a cognitive aid while working with the computer simulation,

while the others only had the instructions. Mean performance was the same in the two groups as was the correlation between hands-on and computer simulation methods: .53. Ruiz-Primo ruled out our first two possible explanations.

Our third possible explanation for the moderate correlation (.53) between hands-on and computer-simulated performance assessment is that the methods tap somewhat different aspects of achievement, as suggested by Shavelson, Baxter, and Pine (1991, p. 360.) If this explanation were tenable, the correlation between Electric Mysteries scores at time 1 and time 2 for either the hands-on or the computer simulation assessment should be higher than the correlation between hands-on scores and computer simulation scores at either point in time. Shavelson, Ruiz-Primo, and Wiley (1999) provided a correlation that did not support this expectation: The correlation between hands-on scores at times 1 and 2 was .57 for observation and for notebook methods. Unfortunately, data were not available to compare the correlation between computer simulation scores at times 1 and 2, the correlation between hands-on scores at time 1 and computer simulation scores at time 2, or the correlation for simulation scores at time 1 and hands-on scores at time 2. Comparing the .53 correlation between simulation and hands-on with the .57 correlation between hands-on at time 1 and time 2, we see that this difference is very small, suggesting both methods may tap the same aspect of achievement.

The sole remaining explanation for the moderate correlation between hands-on and computer simulation scores is that student performance is inconsistent, not because students are unfamiliar with computer simulations or that hands-on and computer simulation methods are so different, but because students are not expert in the subject domain. Different syntactical features of the Electric Mysteries investigation (either hands-on or computer simulation) catch students' attention from one occasion to the next giving rise to inconsistent performance. The roughly constant .53 correlation between the same method at different times is consistent with this explanation.

Student expertise is therefore the most promising explanation for only moderately consistent performance across methods. If this expertise explanation is correct, older, more knowledgeable and experienced students should perform consistently on the hands-on and computer simulation versions of Electric Mysteries than novices. That is, perhaps high school juniors enrolled in an introductory physics course, because of their greater age, exposure to electricity content (having just completed a unit on the topic), and experience with methods of science (i.e.



“experts”) should perform more consistently across occasions and methods. A comparison of this “expert” groups’ performance with the “novice” 5th- and 6th-grade students’ performance would provide a test of the last remaining explanation for only moderate correlations between hands-on and computer simulation methods.

In the expert-novice literature, experts and novices are often differentiated by their ability to see beyond surface features and to recognize similarities in the underlying mechanism of a given problem or situation. Experts tend to have more elaborate cognitive structures within their particular domain of expertise and are better versed in matching the underlying features of a problem situation to these structures. “[M]uch of expert power lies in the expert’s ability to quickly establish correspondence between externally presented events and internal models for these events” (Chi, Feltovich, & Glaser, 1981, p. 123). Because of the expert’s richer and more complete understanding of her particular domain, she is able to see beyond surface features and focus on the mechanisms driving the particular problem or situation. Compared to the novice, who sees surface features as the most salient attributes of a problem (but has only a partial understanding of the underlying features), the expert is not distracted by surface characteristics and sees beyond them. Thus the novice’s understanding centers around declarative knowledge, whereas the expert’s is more procedural and incorporates declarative knowledge as necessary. The “expert’s schemata contain a great deal of procedural knowledge, with explicit conditions for applicability,” whereas the “novice’s schemata may be characterized as containing sufficiently elaborate declarative knowledge about the physical configurations of a potential problem, but lacking abstracted solution methods (Chi, Feltovich, & Glaser, 1981, p. 151).

We reasoned that the elementary students who participated in previous studies had only “partial knowledge” of the design and functioning of basic electric circuits. They were somewhat “novice,” even after having studied about electric circuits in their hands-on science curriculum. Compared to the novices, high school students should have comprehensive knowledge of electrical circuits, having just completed a unit on the topic in their physics course. Moreover, they have had more opportunities during their time in school to perform scientific investigations and problem solve in science, and have had greater exposure to a range of science content. Simply put, we expected that high school students would be more “expert” than elementary students even though high school students may also have partial

knowledge, potentially based on conflicts between the new concepts they are learning and previously constructed knowledge (Mestre, 1994).

## **Method**

### **Subjects and Design**

In the present study, our sample of 40 high school juniors in two physics classes had just completed a unit on electricity. The students attended a public high school in the San Francisco Bay area that had distinguished itself in statewide achievement testing, drawing students from homes where parents were highly educated. The two physics classes were taught by the same teacher and together had a 17 female/23 male ratio. A random half performed the hands-on version of Electric Mysteries first and the computer simulation approximately two weeks later; the other half, vice versa. Sequence of method was balanced for the two classes. In addition to the students' performance on the Electric Mysteries tasks, electricity-unit test scores and overall class grades were provided by the teacher. Assuming these students were "experts," we predicted more consistent performance across methods and occasions and, consequently, higher correlations between scores than found with the 5th and 6th graders.

The data used for the 5th- and 6th-grade students were a subsample of the data used in the original study of the Electric Mysteries investigation by Shavelson, Baxter, and Pine (1991). Our 56 students' scores were taken from a single classroom in Arizona that is known for its exemplary hands-on science program. In the original investigation, these students were tested using both the simulation and hands-on versions of Electric Mysteries separated by one month.

In the current study, these 5th and 6th graders are loosely labeled as "novice" science students due to their limited exposure to science content and processes over their limited time in school. In contrast, we chose high school physics students (juniors and seniors) to be our "experts" because of their more extensive exposure to science in their 13+ years of schooling and because they fall into a more developmentally mature range. Our intent in selecting high school students was to study a group's performance that would potentially provide a strong contrast to the younger students and ultimately support our "expert/novice" hypothesis.

## Electric Mysteries Investigation

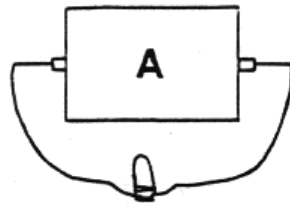
The Electric Mysteries investigation was described at the beginning of the paper. Students recorded their findings in a notebook in which they were asked to (a) record the contents of each mystery box, (b) draw a picture of the circuit they used to determine the contents, and (c) provide a reason for their answer as to the contents (Figure 3).

### Scoring

The original scoring system used by Shavelson, Baxter, and Pine (1991) was based on a scoring strategy in which the student had to correctly identify the contents of a box *and* draw the correct external circuit that was used to identify the contents of the box in order to receive one point (Figure 3). Since either one or zero

Box A: Has a battery and a bulb inside.

Draw a picture of the circuit that told you what was inside BOX A:

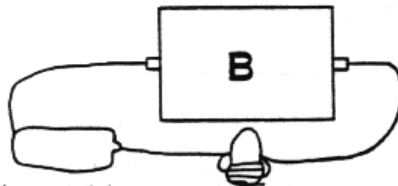


How could you tell from your circuit what was inside BOX A ?

I could tell because the bulb lights, but it was dim so I think there a battery and to light to bulbs.

Box B: Has a wire inside.

Draw a picture of the circuit that told you what was inside BOX B:



How could you tell from your circuit what was inside BOX B?

I think it has just a wire in side because it would not light without a battery.

Figure 3. Electric Mysteries notebook.

points were awarded per box, a maximum of six points was possible for the investigation. The scoring forms for both the hands-on version and the computer simulation version were identical (Figure 4).

Note that Shavelson, Baxter, and Pine (1991) did not incorporate students' justifications into the scoring system. They reasoned that elementary students might be able to perform well yet not be able to explain their performance well in writing. Shavelson et al. did not want to unfairly penalize these students.

For this study we built a new scoring system, one that reflected students' rationales for their responses as well as a more microscopic analysis of score components (Figure 5). We reasoned that expertise differences might very well reside in explanations as well as performance for our "experts." Students were awarded points in four categories:

**Batteries & Bulbs**  
**Electric Mystery Boxes**  
*End of Unit Score Form*

Student \_\_\_\_\_ Scorer \_\_\_\_\_




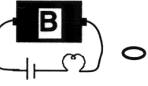

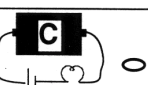

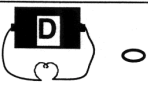



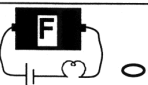
Mystery Box	What's Inside	Circuit	
<b>A</b>	two batteries 		<input type="checkbox"/>
<b>B</b>	bulb 		<input type="checkbox"/>
<b>C</b>	wire 		<input type="checkbox"/>
<b>D</b>	battery and bulb 		<input type="checkbox"/>
<b>E</b>	nothing 		<input type="checkbox"/>
<b>F</b>	wire 		<input type="checkbox"/>
			Total <input type="checkbox"/> 6

Figure 4. Original scoring form.

(note: continue scoring a box only if there is one or two points awarded for the inference)







Box	Inference	Drawing	Observation	Explanation	Score
A	<input type="checkbox"/> 1 Battery, batteries, energy, power inside <input type="checkbox"/> 1 Bulb inside	<input type="checkbox"/> 2 	<input type="checkbox"/> 1 Bulb lights, lights normal, brightly OR <input type="checkbox"/> 2 Bulb lights dimly	<input type="checkbox"/> 1 Battery in box causes voltage/power/energy/brightness <input type="checkbox"/> 1 Bulb in box causes resistance so bulb is dim; discusses a process of resistance taking place or blockage, slowing of electricity	/8
B	<input type="checkbox"/> 0 Nothing OR <input type="checkbox"/> 1 No bulb inside box <input type="checkbox"/> 1 No battery inside box OR <input type="checkbox"/> 2 Wire inside	<input type="checkbox"/> 2 	<input type="checkbox"/> 1 Bulb lights, lights dimly, brightly OR <input type="checkbox"/> 2 Bulb lights with normal brightness	<input type="checkbox"/> 1 Because there is no bulb inside the box, there is no resistance; there is no process of resistance taking place <input type="checkbox"/> 1 Because there is no battery inside the box, there is no voltage/power/energy/brightness OR <input type="checkbox"/> 2 Explicitly states circuit must be completed for bulb to light	/8
C	<input type="checkbox"/> 2 Nothing inside box	<input type="checkbox"/> 2 	<input type="checkbox"/> 2 Bulb does not light, nothing, no response or reaction	<input type="checkbox"/> 2 Explicitly states circuit is not completed or discusses that a circuit must be completed to work	/8
D	<input type="checkbox"/> 1 One battery, energy, battery and bulb inside OR <input type="checkbox"/> 2 Two batteries inside	<input type="checkbox"/> 2 	<input type="checkbox"/> 1 Bulb lights, lights dimly, normal OR <input type="checkbox"/> 2 Bulb lights brightly	<input type="checkbox"/> 2 Battery inside box causes voltage/power/energy to light bulb or circuit lights without an external battery OR <input type="checkbox"/> 2 Two batteries inside box causes extra voltage/power/energy to light bulb brightly – must state some sort of extra power, energy...	/8
E	<input type="checkbox"/> 1 Something is inside, wire inside, battery inside OR <input type="checkbox"/> 2 Bulb inside	<input type="checkbox"/> 2 	<input type="checkbox"/> 1 Bulb lights, lights normal, brightly OR <input type="checkbox"/> 2 Bulb lights dimly	<input type="checkbox"/> 1 What is inside completes the circuit OR <input type="checkbox"/> 2 The bulb inside the box is causing resistance, there is a process of resistance taking place	/8
F	<input type="checkbox"/> 0 Nothing OR <input type="checkbox"/> 1 No bulb inside box <input type="checkbox"/> 1 No battery inside box OR <input type="checkbox"/> 2 Wire inside	<input type="checkbox"/> 2 	<input type="checkbox"/> 1 Bulb lights, lights dimly, brightly OR <input type="checkbox"/> 2 Bulb lights with normal brightness	<input type="checkbox"/> 1 Because there is no bulb inside the box, there is no resistance; there is no process of resistance taking place <input type="checkbox"/> 1 Because there is no battery inside the box, there is no voltage/power/energy/brightness OR <input type="checkbox"/> 2 Explicitly states circuit must be completed for bulb to light	/8
Total (48)					

Figure 5. New scoring form.

- Inference—student’s response as to what is inside the box.
- Drawing—the correct test circuit used to determine the contents of the box.
- Observation—the light bulb’s intensity in reaction to the test circuit and box.
- Explanation—student’s reasoning as to how the Observation related to the Inference.

For each category a maximum of 2 points was awarded, for a maximum of 8 points awarded per box, or 48 points total for the investigation. The awarding of points required the correct statement of a box's contents. If a student scored at least one point in the inference section for a particular box, the other three sections were scored for that box. Otherwise, the student received a zero for that box.

Specifically, our reasoning for developing a new scoring system was based on our conception of a student's ability to make inferences. For example, we felt that if a student was able to determine an electrical component that was *not* in a particular box then she should at least be able to earn partial credit for the inference and be able to earn points in the other three categories. However, if the student did not state any relevant information as to what might or might not be in a box then it did not seem fair to award points in any other category (e.g. providing a correct circuit drawing). The decision to award a greater range of points over a greater number of categories created a scoring form that captured a wider, and potentially more precise, range of student reasoning.

In addition to scoring the correct contents and circuits the students used, the students were scored on the quality of their response to the question "How could you tell from your circuit what was inside box (x)?" We hoped to be able to detect a deeper understanding of a student's content and investigative knowledge through the use of this new scoring rubric. Regarding reliability among raters, both the original scoring rubric and the new scoring rubric had high interrater reliability (.98 for each).

## **Analysis**

The analyses addressed two questions. First, did the high school students ("experts") perform higher than the elementary school students in conducting the Electric Mysteries investigation? Descriptive statistics and analysis of variance were used to address this question, with both the original and new scoring systems. Second, are high school students' scores more consistent across hands-on and computer simulation methods than the scores of elementary school students? Correlations based on both scoring methods were brought to bear on this question.

## **Results**

To help answer our expertise and consistency questions we compared the scores of the high school and elementary school students using both scoring

systems. Addressing our first question, “Did the high school students outperform elementary students in conducting the Electric Mysteries investigation?” analysis of overall scores based on the original scoring form (Figure 4) revealed that the high school students did not perform significantly better than the elementary students on the hands-on method of the investigation (Table 1). However, on the computer simulation we did find a significant difference between the two groups ( $F = 4.615$ ,  $p = .04$ ), with the elementary students scoring higher.

Comparing the elementary and high school student total scores based on the new scoring system<sup>2</sup> we found no significant difference ( $F = .0031$ ,  $p = .9558$ ) between the two groups on the hands-on version of the assessment (Table 2).

We took a closer look at students’ scores using the new scoring form (Figure 5) by dividing student responses into four component scores (inference, drawing, observation, and explanation) of the hands-on task (Table 3). Analyzing our 2 x 4 (grade x component) split-plot design using a repeated-measures ANOVA we found (a) no significant difference by grade ( $F_{1, 94} = 3.96$ ,  $p = .00$ ), (b) a significant difference by component ( $F_{3, 282} = 210.79$ ,  $p = .00$ ), and (c) a significant difference in the component by grade interaction ( $F_{3, 9282} = 8.27$ ,  $p = .00$ ).

Table 1  
Score Summaries Based on Original Scoring System

Grade	Method	Mean	Standard deviation	N
5	Hands-on	3.32	2.09	57
11	Hands-on	2.55	2.26	40
5	Simulation	4.31	1.55	55
11	Simulation	3.50	2.32	40

Table 2  
Score Summaries Based on New Scoring System

Grade	Method	Mean	Standard deviation	N
5	Hands-on	23.84	8.15	56
11	Hands-on	23.95	11.38	40

<sup>2</sup> Detailed scores for the elementary students on the computer simulation are not reported since these scores were not available.

Table 3

Component Score Summaries Based on New Scoring System

Grade	Method	Inference		Drawing		Observation		Explanation	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
5	Hands-on	9.20	2.60	7.39	3.75	6.98	2.76	0.34	1.07
11	Hands-on	9.13	2.91	5.70	4.93	7.15	3.00	1.95	3.14

Both the elementary and the high school students performed similarly on the *inference* and *observation* components. However, for the *drawing* category the elementary student scores were significantly higher ( $F = 3.65, p = .06$ ), whereas for the *explanation* category the high school student scores were significantly higher ( $F = 12.75, p = .0006$ ).

To address our second question, “Are high school students’ scores more consistent across hands-on and computer simulation methods than elementary school students’ scores?” it might be helpful to take a step back and look at the correlation matrix generated from previous studies using the Electric Mysteries investigation with 5th-grade students (Figure 6). From this diagram we can see that the correlation of student performance scores between methods at time 1 ( $r_{H_1C_1}$ ) is almost the same as the correlation for hands-on scores between time 1 and time 2 ( $r_{H_1H_2}$ ). This finding supports our partial-knowledge hypothesis with the 5th-grade students. That is, since the 5th-grade students only have partial knowledge of the

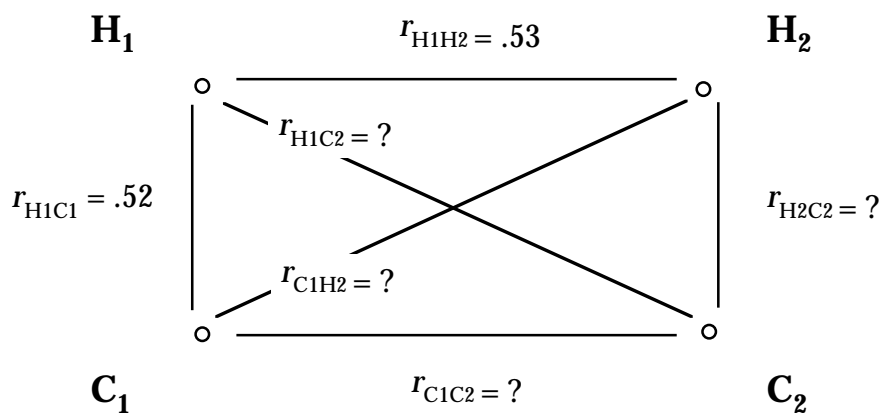


Figure 6. Known and unknown correlations between hands-on and computer simulation performance scores taken at two points in time based on previous Electric Mysteries studies with 5th-grade students.



procedures and content required to perform the Electric Mysteries investigation, the correlation of their performance scores will be similar across methods and across time. Comparing the .53 correlation between simulation and hands-on with the .52 correlation between hands-on at time 1 and time 2, we see that this difference is very small, suggesting both methods tap the same aspect of achievement.

The current study with the 11th-grade physics students addressed part of our partial-knowledge hypothesis by looking at their cross-lagged correlations ( $r_{H1C2}$  and  $r_{C1H2}$ ). As we stated earlier, it may be that issues of exchangeability are not due to method, but rather to students' partial knowledge within the specific science domain. In other words, if the partial-knowledge explanation is tenable (that performance is inconsistent not from one method to another, but from one occasion to another), all correlations, regardless of method or occasion would be of the same magnitude (about .53). Analyzing the cross-lagged correlations we found that  $r_{H1C2} = .64$  ( $n = 19$ ,  $p = .003$ ) and  $r_{H2C1} = .40$  ( $n = 21$ ,  $p = .074$ ), the average of which puts us at .52, almost identical to the .53 correlation suggested above. Unfortunately, from the design of the current study we were unable to determine the correlations  $r_{H1C1}$ ,  $r_{H1H2}$ ,  $r_{C1C2}$ , or  $r_{H2C2}$ , which would provide a more complete answer to our partial-knowledge hypothesis. We leave the determination of these additional correlations to further studies.

With regard to the second question, even though we did not find the expected novice/expert difference, we nevertheless compared hands-on and computer simulation correlations for elementary students ( $n = 55$ ) with the same correlations for high school students ( $n = 40$ ). Using the old scoring rubric, we found a considerably larger difference between assessment methods ( $r = .36$ ,  $p = .007$ ) than reported earlier ( $r = .53$ ). For the 11th-grade students, the difference between methods was even greater ( $r = .21$ ,  $p = .20$ ).

## Discussion

The comparisons between groups using the original scoring system showed that the elementary students outperformed the high school students on both the hands-on and computer simulation methods. Part of this difference was due to missing or incomplete drawings by the high school students. In our scoring of the assessments, the high school students were more likely to omit the drawing section than were the elementary students, even though both groups were instructed to

complete the drawings.<sup>3</sup> The reason for this is not clear. Therefore, based on this information the suggestion that the high school students were still “novice” in their thinking about the underlying mechanisms of the investigation needs to be taken with caution—the results are inconclusive.

Similar comparisons were made between the two groups using the new scoring system. Since the design of the new scoring system was based on how we thought of a student’s ability to make inferences, we expected to be able to detect a deeper understanding of content and investigative knowledge regarding electric circuits. Here we found that the high school and the elementary student scores differed on several of the score components. Even though the high school students had lower overall mean scores for the hands-on method, they tended to provide better explanations linking their observations and inferences. For instance, the high school students provided better concept-based reasons for why the light bulb responded the way it did to a given test circuit, and used the correct terminology (such as “resistance” or “completion of circuit”) more often to describe what was happening in the investigation than did elementary students.

The higher mean “explanation” score of the high school group may suggest that high school students have deeper declarative knowledge of electric circuits than the elementary students do, or are better able to verbalize this knowledge. Based on their ability to provide more salient reasons for their inferences, even though their “inference” scores were almost identical to the elementary students’ scores, high school students may in fact have a better sense of the underlying mechanisms. However, it is also possible that the high school students have a wider range of declarative knowledge than the elementary students and occasionally match the right terminology with the right procedure to produce a better “explanation.” Thus, the high school students still may be focusing on the surface features of a problem, but they are better able to provide a richer description of what they think is happening.

If the high school students are in fact more “expert,” they should also show high performance consistency between method and occasion—the investigation should not look considerably different because it is in a different form (hands-on or computer simulation) or because there is a delay between the two testing occasions.

---

<sup>3</sup> For example, on the hands-on version 14 high school students out of 40 omitted the drawing, whereas only 1 of 56 elementary students left the drawing blank.

However, this was not the case. The correlation between methods (delayed by approximately two weeks) for the high school students was .21, which is considerably lower than the original correlation of .53 reported by Shavelson, Baxter, and Pine (1991), delayed by one month. Because both occasion and method varied, performance of students with partial knowledge varied considerably.

The results of our study suggest that high school students are not necessarily “expert” when compared with the elementary students on the hands-on and computer simulation versions of the Electric Mysteries assessment. Although more verbose in their descriptions, high school students were still primarily attracted to the surface features of a given problem. Because they had partial knowledge of the physical principles underlying electric circuits, their scores varied considerably from one method to another and from one occasion to the next.

Going back to our original questions—Are the hands-on and computer versions the same? Are scores earned in one medium the same as (“exchangeable for”) scores earned in the other?—our study does not provide definitive answers. We chose high school physics students to be our experts, expecting them to perform consistently across methods and occasions, but at the same time be able to show us if the hands-on and computer simulation versions tap different knowledge. However, since our evidence suggests that these high school students may not be more “expert” than elementary students (scores varied considerably between methods and occasions based on their partial knowledge), the exchangeability of methods question remains unanswered. In other words, the issue of exchangeability is still confounded with inconsistencies in the high school students’ performance.

Further investigations may provide more definitive answers. One investigation, mentioned previously, might be to test high school students twice, either on the same method or both methods on the same day or on two days separated by a few weeks. In other words,  $r_{H1C1}$ ,  $r_{H1H2}$ ,  $r_{C1C2}$ , and  $r_{H2C2}$  would be determined to provide a more complete answer to our partial-knowledge hypothesis.

A second investigation would combine the above design with more “expert” students, perhaps undergraduate physics majors, first-year graduate physics students, or high school physics teachers. Again, this further investigation may provide a more definitive answer to the issue of exchangeability of the hands-on and computer simulation versions of the Electric Mysteries assessment.

## References

- Baxter, G. P. (1995). Using computer simulations to assess hands-on science learning. *Journal of Science Education and Technology, 4*(1), 21-27.
- Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research, 21*, 279-298.
- Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*(3), 5-12.
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 6*, 37-75.
- Mestre, J. P. (1994). Cognitive aspects of learning and teaching science. In S. J. Fitzsimmons & L. C. Kerpelman (Eds.), *Teacher enhancement for elementary and secondary science and mathematics: Status, issues, and problems* (pp. 3-1-3-53; NSF 94-80). Washington, DC: National Science Foundation.
- Pine, J., Baxter, G. P., & Shavelson, R. J. (1993). Assessments for hands-on elementary science curricula. *MSTA Journal, 39*(2), 3, 5-19.
- Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*, 347-362.
- Shavelson, R. J., Baxter, G. P., Pine, J., & Yure, J. (1991, April). *Hands-on performance assessments and their notebook surrogates*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- 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.
- Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice, 16*, 16-25.
- Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19*, 1-14.