On the Stability of Performance Assessments

Maria Araceli Ruiz-Primo, Gail P. Baxter, and Richard J. Shavelson
University of California, Santa Barbara

This study examined the stability of scores on two types of performance assessments, an observed hands-on investigation and a notebook surrogate. Twenty-nine sixth-grade students in a hands-on inquiry-based science curriculum completed three investigations on two occasions separated by 5 months. Results indicated that: (a) the generalizability across occasions for relative decisions was, on average, moderate for the observed investigations (.52) and the notebooks (.50); (b) the generalizability for absolute decisions was only slightly lower; (c) the major source of measurement error was the person by occasion (residual) interaction; and (d) the procedures students used to carry out the investigations tended to change from one occasion to the other.

Cognitive research, curriculum reform, and limitations of multiple-choice testing have all motivated the search for alternative methods for assessing science achievement (e.g., Frederiksen, 1984; Frederiksen & Collins, 1989; Linn, 1986; Nickerson, 1989; Shavelson, Baxter, Pine, Yuré, Goldman, & Smith, 1991c; Shavelson, Baxter, Pine, & Yuré, 1991b). One alternative to multiple-choice testing, congruent with curricular reform and constructivist learning theories, is hands-on performance assessments. In science, for example, a performance assessment provides students with laboratory equipment, poses a problem, and allows them to use these resources to generate a solution. This hands-on test is administered on a one-to-one basis, and performance is judged in real time as experts watch students do science. Scores for these hands-on assessments reflect not only the adequacy of students’ solutions but also the procedures used to arrive at their solution (Baxter, Shavelson, Goldman, & Pine, 1992).

Observed hands-on assessment is not practical on a large-scale basis due to costs of equipment, personnel, and testing time. In practice, students typically conduct an investigation and record, in a notebook, their experimental procedures, results, and conclusions (Tamir, 1974). These notebooks can then be scored in roughly the same manner as the actual performance (e.g., Baxter et al., 1992; Shavelson, Baxter, Pine, & Yuré, 1991b).

Shavelson, Baxter, and Pine (1991a) developed and examined the psychometric properties of three observed hands-on science investigations, and their corresponding notebooks: paper towels (determine which of three paper towels absorbs the most and least water), electric mysteries (determine the circuit components hidden in six black boxes), and bugs (determine the preferences of sow bugs for various environments).

Over a 2-year period more than 300 fifth- and sixth-grade students, some experienced and some inexperienced in hands-on science curriculum, were administered a battery of performance assessments, a traditional multiple-
choice science achievement test (Comprehensive Test of Basic Skills in Science, CTBS; CTB/McGraw-Hill, 1989), and a cognitive abilities test (CAT; Shavelson et al., 1991b). All these data were brought to bear on issues of reliability and validity. During the first year, two raters scored the observed hands-on performance and notebook measures. During the second year, based on the results of the former study, only one rater scored students' performance. Nevertheless, the performance of a sample of students (N between 10 and 20) was evaluated by a second observer (shadow), and interrater reliability was estimated on this sample.

Results were consistent over the 2 years. For the observed hands-on investigations: (a) Interrater reliability was consistently high for all investigations (> .90) and varied little by the curricular experience of the students; (b) interrater agreement was high on procedures students used to conduct the investigation (> .90); (c) a large person x task interaction for the electric mysteries investigation suggested that more boxes were needed to attain a consistent measure of student performance (adding one more box in the second year to the five already used in the first year led to an increase in reliability from .65 to .73); (d) for all investigations, the mean performance of students experienced in hands-on science was higher than that of the students not experienced (the largest mean difference was on the investigation that required domain specific knowledge—electric mysteries); (e) the correlation between scores for observed investigations and scores on a traditional multiple-choice science test was only moderate in magnitude (mean r = .42) suggesting, along with other analyses (e.g., Shavelson et al., 1991b), that hands-on assessments may measure different aspects of science achievement;¹ and (f) scores on observed investigations correlated lower with aptitude than did the traditional science achievement test (mean r = .37 and .59, respectively).

For the notebook, interrater reliability was slightly less than that for observed investigations (mean = .88), and interrater agreement on procedures was lower (mean = .79). Mean performance was about the same, and the correlations with aptitude and the traditional science achievement test were of the same magnitude as those for the observed investigation (Shavelson et al., 1991b).

The present study extends the Shavelson, Baxter, and Pine (1991a) research on the psychometric characteristics of performance assessments and their less costly surrogates by examining the stability (test–retest reliability) of performance assessments—observed hands-on investigations and their notebook surrogates. Even though students are typically tested on one occasion, educators, policymakers, and parents generalize performance scores across occasions. The question is, if students were tested on another day, would the estimate of their science achievement change? An investigation of the stability of performance measures adds to understanding how well the performance assessments describe students’ achievement. Since stability studies are costly and time consuming, the study of stability of performance assessments is rarely considered, despite its importance.

¹ See the note at the end of the text.
This study examined stability of performance assessments within the context of generalizability (G) theory. G theory recognizes that multiple sources of error contribute to the unreliability of a measure and hence the estimate of student performance (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Johnson & Bell, 1985; Shavelson, Webb, & Rowley, 1989). G theory uses the analysis of variance to provide estimates of performance variation due to persons and each source of error. By estimating the magnitude of the variance components, the source(s) of the greatest measurement error can be pinpointed (e.g., tasks, occasions, or their interactions). Moreover, G theory provides an estimate of the reliability (generalizability) coefficient for two types of interpretations: relative and absolute. Relative interpretations focus on the rank ordering of students, as is common in correlational analyses. Absolute interpretations focus on the level of performance in a domain without regard to the performance of others (Shavelson & Webb, 1991).

We asked in this study the following specific questions: First, is performance generalizable across occasions for the observed investigations and the notebook surrogates? Tamir (1974) measured science performance using equivalent problems and different raters on two occasions across 16 schools. The mean reliability coefficient for this design was .35 with a range of .12 to .85 across the schools. Previous research in another domain—military job performance—has shown retest reliabilities on the order of .70. Moreover, this research has shown that mean performance scores increased from the first to the second test occasion (Carey, 1991; Mayberry & Hiatt, 1990). Can the same be expected of elementary students’ performance in science?

Second, is the estimate of students’ achievement, based on all three investigations, stable across the two occasions? Previous research suggests unreliability due to task sampling (Shavelson & Baxter, 1992; Shavelson, Baxter, & Pine, 1991a). Since a large number of tasks is needed to reliably estimate student performance, we expected that combining the three investigations into a total score would increase the stability coefficient.

Third, do students conduct the investigations in the same way on each occasion? In other words, are the same sequences of procedures used each time? Siegler (1987, 1989) and Siegler and McGilly (1989) found, on a wide variety of mathematical tasks, that children frequently used different strategies to solve the same problem on two different occasions. We expected students to change their problem-solving strategies in carrying out the investigations and held out the possibility that their overall level of performance might increase, even though no further instruction or practice was received.

Method

Subjects

Twenty-nine students, experienced in hands-on science, were randomly selected from a larger sample (N = 110) from a previous study (Shavelson et al., 1991b). Because the concern of the present study was the stability of measures across occasions, students were selected following two criteria: (a) They were enrolled as regular seventh-grade students in the junior high school
in which this research was conducted, and (b) they had been evaluated previously by one of the four observers participating on the second occasion. Although 29 students were tested, not all of them completed the three investigations \(n_{\text{paper towels}} = 29; n_{\text{electric mysteries}} = 28; n_{\text{bugs}} = 27\).

**Instrumentation**

**Observed investigations.** Students completed three investigations. The first, paper towels, asked them to determine which of three paper towels soaked up the most water and which soaked up the least water. The laboratory equipment, which included a pitcher of water, trays, beakers, a scale, a timer, eyedropper, and so forth, allowed the investigation to be conducted with a variety of solution strategies.

The second investigation, electric mysteries, asked students to determine the circuit components hidden in six mystery boxes. Given a set of five wires, two batteries, and two bulbs, students were asked to determine the contents of a given box from a list of five possible alternatives (e.g., one alternative is two batteries, another is a battery and a bulb). Two boxes had the same contents.

The third investigation, bugs, asked students to determine sow bugs’ preferences for various environments. Students conducted three experiments: The first varied dark or light areas, the second varied wet or dry areas, and the third varied the \(2 \times 2\) factorial combination of the first two experiments. The students were provided with bugs, a lamp (to create a light condition), dishes, black paper (to create a dark condition), scissors, and a spray bottle (to create a damp condition).

**Notebooks.** Students were asked to complete scientific notebooks wherein they described their investigations such that a friend could replicate them. For the paper towels investigation, students were asked to record the steps in the investigation, the variables controlled, and the basis on which they arrived at their conclusions. For the electric mysteries investigation, students were asked to: (a) note the contents of each box, (b) draw a diagram of the circuit used to determine the contents, and (c) write a reason for the decision as to the contents of a particular box. For each of the three bugs experiments, students were asked to: (a) draw a picture depicting the number of bugs they used and how they set up the experiment (e.g., one or two conditions in Experiments 1 and 2; one, two, or four conditions in Experiment 3), (b) describe the steps in the experiment, and (c) indicate how they arrived at their conclusions.

**Scoring**

A procedure-based scoring system was used to score the paper towels investigation. Observers identified the method students’ used to wet the towels, the degree of saturation, the measurement taken to determine the result, and the correctness of the solution generated. Each combination of method, saturation, and measurement produced a unique procedure. A total of 36 possible procedures could be coded; 16 were observed (Baxter et al., 1992). A letter grade was assigned based on the scientific soundness of the observed
procedure (from 1 = F to 14 = A+). In this way, several equally adequate procedures could result in the same letter grade.

The scoring system used for the electric mysteries investigation focused on the correctness of the answer and sequence of circuits used to arrive at the answer. Observers recorded the sequence of circuits used by the students to determine the contents of each box. For example, a student might hook up a bulb to the box. When the bulb did not light, he or she may then have put a battery in the external circuit. Many sequences were possible (108 in all); they were coded into five categories: (a) single circuit that is the correct circuit to discover the contents of the box, (b) single circuit that is not the correct circuit for the box, (c) a correct sequence of two circuits in which the student first connects a bulb and then connects a battery and bulb to determine the contents of a box that does not contain a battery, (d) a sequence of more than two circuits that contains the correct sequence (e.g., bulb, battery and bulb, bulb again to make sure), and (e) a sequence of two or more circuits that does not include the correct circuit for determining the contents of the box. To receive a score of 1 on each box, students had not only to correctly identify the contents but also to use the correct sequence of circuits. The total score could range from 0 to 6.

The bugs scoring system was procedure-based like that used for paper towels. For each experiment, observers focused on the method used to solve the problem (e.g., two conditions in a dish or one condition at a time), the adequacy with which conditions were manipulated (e.g., equal area for each condition), the measurement strategies used to determine the result (e.g., count the number of bugs in each condition), and the correctness of the solution generated. As with paper towels, observers noted the procedure used and assigned a letter grade for each experiment (1 = F to 14 = A+). Like the paper towels investigation, several different procedures could result in the same letter grade.

The notebooks were evaluated with the same scoring system used for the corresponding observed investigation. For the paper towels and bugs investigations, sequences of procedures were noted, and a letter grade was assigned. For electric mysteries notebooks, students recorded the circuit used to determine the contents of the box. Hence, scores were based on what the students said was in the box and the circuit reported in making this determination (in contrast to a sequence of circuits for the observed investigation).

For convenience in analyzing and reporting the data, the 14-point scores for paper towels and bugs were converted to scales ranging from 0 to a maximum of 6 like those used in scoring electric mysteries. Specifically, paper towels and bugs scores were divided by a constant, 2.33.

Training

For each investigation, observers were trained as follows: All observers carried out the investigations and discussed their procedures. Next, each component of the scoring system was presented. Observers then watched videotapes of students conducting the investigations. Scoring was discussed to
clarify any misunderstandings. Then observers independently scored five videotapes. Interrater reliability was typically quite high, above .90.

Procedure

Students were tested on two occasions separated by 5 months. During this period (May–October), students did not receive instruction directly related to any of the investigations. On the first occasion, students were randomly assigned to one of eight observers. Four of these eight observers participated on the second occasion. Twenty-two of the 29 students were tested by the same observer on both occasions. For the remaining 7 students, a different observer was used on each occasion. (Previous research has shown that this procedure does not pose problems; see Shavelson, Mayberry, Li, & Webb, 1990).

On both occasions, each student was observed conducting the investigations in the same order: paper towels, electric mysteries, and bugs. For each investigation, instructions were read aloud to each student, and each piece of equipment was introduced. For both paper towels and bugs, the observer noted the sequence of procedures used and assigned a letter grade while the student performed the investigation. For the electric mysteries investigation, observers recorded the sequence of circuits students used to determine the contents of each box and the students’ conclusions.

Following the paper towels and the bugs investigations, students completed notebooks in such a way that a friend could replicate the investigation. The electric mysteries notebooks were completed as students conducted the investigation. All notebooks were scored by the same rater approximately 3 months later using the same scoring system as for the observed investigations.

Results and Discussion

This study addressed three questions: First, is performance generalizable across occasions for the observed investigations and the notebook surrogates? Second, is the estimate of students’ achievement, based on all three investigations, stable across occasions? And third, do students conduct the investigations in the same way on each occasion?

Generalizability of Scores

To examine the generalizability of scores, person × occasion G studies were carried out for each investigation and method. For both observed investigations and their notebook surrogates, G coefficients were roughly of the same magnitude (Table 1); the average over the three investigations was 0.52 and 0.50, respectively.

The low generalizability was due largely to the residual (p × o, e) variance component (Table 1). Students’ relative standing varied substantially from one occasion to the other. Increasing the number of occasions would increase the coefficients (four occasions to achieve reliability .80), but it would do so at considerable cost.

The G coefficients for relative decisions were slightly higher than the absolute decisions. The difference is due to sampling fluctuation in mean
## Stability

### TABLE 1
Estimated Variance Components and Coefficients for all Observed Investigations and Notebook Surrogates

<table>
<thead>
<tr>
<th>Investigation</th>
<th>Source of Variation</th>
<th>Method</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Observed</td>
<td>Notebooks</td>
<td></td>
</tr>
<tr>
<td>Paper Towels</td>
<td>Persons (p)</td>
<td>0.81</td>
<td>1.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occasion (o)</td>
<td>0.12</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p x o, e</td>
<td>0.79</td>
<td>1.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p^2</td>
<td>.50</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \phi )</td>
<td>.46</td>
<td>.50</td>
<td></td>
</tr>
<tr>
<td>Electric Mysteries</td>
<td>Persons (p)</td>
<td>1.70</td>
<td>1.69</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occasion (o)</td>
<td>0.26</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p x o, e</td>
<td>1.30</td>
<td>1.62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p^2</td>
<td>.57</td>
<td>.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \phi )</td>
<td>.52</td>
<td>.45</td>
<td></td>
</tr>
<tr>
<td>Bugs</td>
<td>Persons (p)</td>
<td>0.63</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occasion (o)</td>
<td>0.00</td>
<td>0^a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p x o, e</td>
<td>0.65</td>
<td>0.47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p^2</td>
<td>.49</td>
<td>.49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \phi )</td>
<td>.48</td>
<td>.49</td>
<td></td>
</tr>
</tbody>
</table>

^a negative variance components set to zero (-0.001).

performance from one occasion to the other. Although all three means for investigations and methods (observed and notebook surrogate) were higher on the second occasion (Table 2), these differences were not statistically significant for either observed or notebook data (MANOVA: \( F_{2,51} = .569, p > .05; F_{2,51} = 1.246, p > .05 \)).

### Generalizability of Total Scores

Recognizing that generalizations will be drawn across all three investigations to form an estimate of students’ science achievement, we examined the stability of a combined (total) score in a person \( \times \) occasion G study (Table 3). As expected, the generalizability coefficients for both the observed investigations and the notebook surrogates were considerably higher than coefficients reported for any single investigation, suggesting that the students’ science achievement was reasonably stable across occasions. The relative magnitude of
Ruiz-Primo, Baxter, and Shavelson

TABLE 2
Comparison of Mean Scores Across the Two Occasions

<table>
<thead>
<tr>
<th>Investigation</th>
<th>Method: Observed Mean S.D.</th>
<th>Notebooks Mean S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Maximum Score = (6)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Occasion</td>
<td></td>
</tr>
<tr>
<td>Paper Towels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.89 1.42</td>
<td>4.39 1.64</td>
</tr>
<tr>
<td>2</td>
<td>4.36 1.43</td>
<td>4.70 1.43</td>
</tr>
<tr>
<td>Electric Mysteries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.93 1.88</td>
<td>3.34 2.00</td>
</tr>
<tr>
<td>2</td>
<td>4.67 1.54</td>
<td>4.17 1.65</td>
</tr>
<tr>
<td>Bugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.26 0.98</td>
<td>4.42 0.96</td>
</tr>
<tr>
<td>2</td>
<td>4.51 1.27</td>
<td>4.60 0.96</td>
</tr>
</tbody>
</table>

the residual variance component \((p \times o, e)\) was reduced substantially by combining investigations. From a practical perspective, generalizability of 0.80 can be achieved with two occasions (based on the G theory analog to the Spearman-Brown prophecy formula).

Stability of Experimental Procedures

Since students’ relative performances varied across occasions, we expected to see variations in their procedures across occasions. Thus, we examined agreement in procedures by comparing procedural sequence codes on the two occasions. Agreement was defined as the number of exact code matches between the two occasions divided by the total possible matches. Note that students could change procedures across occasions but their performance score might stay the same because several different procedures might result in

TABLE 3
Estimated Variance Components and Coefficients for Total Scores of All Observed Investigations and Notebooks Surrogates with a Person x Occasion G Study

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Method: Observed</th>
<th>Notebooks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons (p)</td>
<td>0.74</td>
<td>0.61</td>
</tr>
<tr>
<td>Occasion (o)</td>
<td>0.11</td>
<td>0.09</td>
</tr>
<tr>
<td>(p \times o, e)</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>(\rho^2)</td>
<td>.73</td>
<td>.68</td>
</tr>
<tr>
<td>(\phi)</td>
<td>.65</td>
<td>.62</td>
</tr>
</tbody>
</table>
the same letter grade. For all the investigations, the low to moderate agreement coefficients indicated that students used different procedures on each of the two occasions (Table 4).

*Paper towels.* For the paper towels investigation, students usually changed only one component of their procedures; changes were not haphazard. Indeed, only two students used a completely different procedure—different in the method used for wetting the towel, different in the degree of saturation, and different in the method used to determine the result—on each occasion.

In general, procedural agreement was higher for notebooks than observed investigations. In the notebooks, students reported procedures in a general way, and the rater often had difficulty determining what they did from what they wrote (Baxter et al., 1992). To score notebooks, several rules were developed. For example, if students did not explicitly say that they saturated the towels but checked “Yes” to the question; “Were the towels completely wet?”, credit was given for saturation. However, in reality, students may have controlled saturation on Occasion 1 and saturated the towel on Occasion 2, or vice versa. This would result in agreement of procedures for notebooks but not for observed performance.

*Electric mysteries.* The percent agreement for the electric mysteries investigation was considerably higher than that for the other two investigations. Percentage of agreement varied across mystery boxes. Disagreements across occasions arose for at least two reasons: (a) More students used the correct circuits on the second occasion, which decreased the agreements and increased the number of correct answers; (b) students used more circuits on Occasion 1 than on Occasion 2, giving the impression of a trial-and-error approach on the first occasion and a more systematic approach, due to familiarity, on the second.

<table>
<thead>
<tr>
<th>Investigation</th>
<th>Method: Observed Percentage</th>
<th>Method: Notebook Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper Towels</td>
<td>24</td>
<td>27</td>
</tr>
<tr>
<td>Electric Mysteries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Box A</td>
<td>53</td>
<td>67</td>
</tr>
<tr>
<td>Box B</td>
<td>60</td>
<td>67</td>
</tr>
<tr>
<td>Box C</td>
<td>55</td>
<td>64</td>
</tr>
<tr>
<td>Box D</td>
<td>59</td>
<td>64</td>
</tr>
<tr>
<td>Box E</td>
<td>44</td>
<td>60</td>
</tr>
<tr>
<td>Box F</td>
<td>44</td>
<td>46</td>
</tr>
<tr>
<td>Bugs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bugs 1</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>Bugs 2</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>Bugs 3</td>
<td>29</td>
<td>40</td>
</tr>
</tbody>
</table>

49
Procedural agreement for electric mysteries notebooks was higher than for observed performance. This difference can be attributed to assessment procedure variation. In the notebooks, students recorded the one circuit that led them to their decision about the contents of the box. In contrast, observers recorded the sequence of every circuit that the students constructed to determine the contents of a box. Hence, there was less chance for disagreement.

**Bugs.** Procedural agreement was calculated for each of the three observed bugs experiments (Table 4). As with paper towels, the coefficients were low. Again, the disagreement did not reflect totally different procedures. Rather, students might have changed one or two components (e.g., number of bugs and/or time of observation). Nevertheless, the pattern of agreement was curious. Why was agreement on procedures for Experiments 1 and 3 lower than that for Experiment 2? One possible explanation may be that Experiment 1 served as a training task for Experiment 2 on both occasions. For example, students tended to use more bugs for a longer time in Experiment 2 than in Experiment 1—thus, the higher percentage of agreement in Experiment 2 (Table 4). However, Experiment 3 was more difficult because a factorial design was required. Even though the increase in bugs and time carried over, the students appeared to be searching for an appropriate experimental design.

For the bugs notebooks, procedural agreement was calculated for each experiment (Table 4). Disagreements arose largely over the method used to determine the result. As with the paper towels, the information given by students in their notebooks was not always consistent with the information reported by the observer. For example, observers noted that students counted the number of bugs in each area on both occasions. According to what students wrote in their notebooks, on Occasion 1 they took into account the time the bugs stayed in an area while on Occasion 2 they counted the number of bugs in each area. This resulted in greater consistency for observed performance than for notebooks. However, when we considered as equivalent the bug count in each area and the time they stayed in an area, the agreement coefficients increased to .48 for percent agreement in Experiment 1 and .62 (respectively) in Experiment 2.

In general, the procedural agreement data across occasions was consistent with strategy use research in the cognitive psychology literature. When a problem (i.e., subtraction, multiplication) was presented to the same children twice, they frequently used different strategies. Some children changed from a more advanced to a less advanced strategy, others did the reverse (Siegler, 1989).

**Conclusions**

This study examined the generalizability of performance assessment scores and procedures across occasions. Three questions were raised: Is performance generalizable across occasions for the observed investigations and the notebook surrogates? Is the estimate of students’ achievement, based on all three investigations, stable across the two occasions? And, do students conduct the investigations in the same way on each occasion?
Stability

Our findings led to the following tentative conclusions for all three investigations: (a) Students’ performances changed from one occasion to the next—G coefficients for absolute decisions were consistent for observed investigations and notebooks—on average, .48. (b) When students’ scores on the individual investigations were aggregated to produce a science achievement score, generalizability for absolute decisions increased substantially (.65 and .62 for observed performance and notebooks, respectively). (c) The procedures students used to conduct the investigations changed from one occasion to the next. In general, their performance was more focused on Occasion 2 than on Occasion 1.

Inevitably educators and the public will generalize a student's performance score, earned on one occasion, to a whole set of possible occasions (e.g., in fifth grade). To increase the dependability of this generalization, a substantial number of investigations or more occasions will need to be sampled. Either strategy for increasing reliability will be costly and time consuming. Nevertheless, the stakes in such large-scale assessment seem to warrant the expense.

Notes

1This interpretation is based on additional analyses including disattenuated correlations and multiple correlation between the multiple-choice test and a combination of performance assessments in life and physical science. Nevertheless, counterinterpretations should not be ruled out.

2Dollars, time, space, and access costs associated with tracking students from elementary to junior high school and with collecting retest data on the performance assessments were extraordinarily high. Hence the small sample of 29 subjects.

3The reader is referred to Baxter, Shavelson, Goldman, and Pine (1992) for complete details of the procedure-based scoring system and the training of the observers.

4A person nested within rater design with both persons and raters crossed with occasions (p: r x o) could have been used to examine these data. However, raters were so well calibrated that introducing this complexity seemed unnecessary. Therefore, we assume that raters are exchangeable (Shavelson et al., 1991).

5For the paper towels investigation, one student received an A in the observed investigation and a D in the notebook while a second student did just the opposite. These two outliers were eliminated from the analysis because they overinfluenced the pattern of results.

References


1153–1160.
searcher, 18(9), 3–7.
science. Educational Leadership, 49(8), 20–25.
assessments and their notebook surrogates. Paper presented at the Annual Meeting of
the American Educational Research Association, Chicago.
Alternative technologies of large-scale science assessment: Instruments of education
Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achieve-
ment: Options for a powerful policy instrument. Phi Delta Kappan, 71(9), 692–697.
CA: Sage.
4–9.
searcher, 18(9), 15–20.
Levin & D. Zakay (Eds.), Time and human cognition: A life-span perspective (pp.
Educational Researcher, 20, 9–11.
Measurement, 11, 25–33.
41–47.
Delta Kappan, 70(9), 703–713.

Authors

MARIA ARACELI RUIZ-PRIMO is Assistant Professor, National University of
Mexico, and PhD Candidate, Graduate School of Education, Phelps Hall, University
of California, Santa Barbara, CA 93106. Degrees: BA, MA, National University of
Mexico. Specializations: measurement and program evaluation.
GAIL P. BAXTER is Assistant Professor, University of Michigan, 610 E. University, SEB 4115, Ann Arbor, MI 48109. Degree: PhD, University of California–Santa Barbara. Specialization: educational measurement.

RICHARD J. SHAVELSON is Dean and Professor, Graduate School of Education, Phelps Hall, University of California, Santa Barbara, CA 93106. Degrees: BA, University of Oregon; MA, San Jose State University; PhD, Stanford University. Specializations: measurement, education, and human resource policy analysis.