Reasoning Dimensions Underlying Science Achievement: The Case of Performance Assessment

Carlos Cuauhtémoc Ayala, Richard J. Shavelson, and Yue Yin
School of Education
Stanford University/CRESST

Susan E. Schultz
School of Education
Stanford University

Snow argued for multidimensional science achievement in the National Education Longitudinal Study of 1988 (NELS:88) along dimensions of basic knowledge and reasoning, spatial–mechanical reasoning, and quantitative science. We focused the generality of these reasoning dimensions in other multiple-choice tests and performance assessments. Confirmatory factor analyses retrieved the 3 dimensions for a test composed of NELS:88, the Third International Mathematics and Science Study (TIMSS) and the National Assessment of Educational Progress (NAEP) multiple-choice items, and the NELS:88 items alone. We used the latter because factor correlations were lower. We administered 3 reasoning-dimension-linked performance assessments to a subsample of 35 students from the main study. Performance assessments correlated moderately with each other and NELS:88 reasoning scores; the 2 methods partially converged on the dimensions. Performance scores scattered across multiple-choice scores due to the broad reasoning and knowledge spectrum tapped. Findings are tentative; larger samples and cognitive studies of reasoning and knowledge might shed light on convergence.

This study sought to determine whether the reasoning dimensions identified by Snow and colleagues (Hamilton, Nussbaum, & Snow, 1997)—basic knowledge

Requests for reprints should be sent to Richard J. Shavelson, School of Education, Cubberley Hall, 485 Lasuen Mall, Stanford University, Stanford, CA 94305.
and reasoning, spatial–mechanical reasoning, and quantitative science—were unique to NELS:88 multiple-choice test or could be found not only in other multiple-choice tests, but also in alternative assessments such as performance assessments. To this end, we built and analyzed a multiple-choice science achievement test composed of National Education Longitudinal Study of 1988 (NELS:88), the Third International Mathematics and Science Study (TIMSS), and National Assessment of Educational Progress (NAEP) items. We also selected three performance assessments—Electric Mysteries, Daytime Astronomy, and Aquacraft—to tap, among other things, the three reasoning dimensions. We fully recognized that performance assessments were far more complex than multiple-choice items and were likely to measure other aspects of science achievement as well, for example, declarative ("knowing that"), procedural ("knowing how"), and schematic ("knowing why") knowledge (Li & Shavelson, 2001; Li, Shavelson, & White, 2002; Shavelson & Ruiz-Primo, 1999). Nevertheless, if the reasoning dimensions were present in other assessment methods, this would increase confidence in interpretations of science achievement test scores, as reflecting in part, these dimensions.

**REASONING DIMENSIONS**

Hamilton et al. (1997) found three reasoning dimensions underlying students’ science achievement on the NELS:88 multiple-choice test—basic knowledge and reasoning (BKR), quantitative science (QS), and spatial–mechanical reasoning (SM)—that were confirmed with small-scale interviews (Hamilton et al., 1997; Nussbaum, Hamilton, & Snow, 1997; see Shavelson et al., this issue). Basic knowledge and reasoning draws on general verbal reasoning using declarative (facts, concepts) knowledge. Quantitative science reasoning involves the manipulation of numerical quantities and requires specialized, classroom-based knowledge. And spatial–mechanical reasoning requires reasoning about visual or spatial relations, motions, distances, or some combination of these. Corroborating evidence supporting the three reasoning dimensions came from think-aloud protocols, observations, and posttest interviews (Hamilton et al., 1997). Furthermore, Hamilton and Snow (1998) identified salient features of multiple-choice and constructed response items that revealed the largest difference in scores. For example, the spatial–mechanical dimension, which revealed a gender effect, could be differentiated from the other reasoning dimensions based on students’ more frequent use of predictions, gestures, and visualization.

We posited three reasoning dimensions following Snow’s earlier work and set out to determine whether or not these reasoning dimensions might be present in multiple-choice tests other than NELS:88 and assessments other than multiple-choice tests, such as performance assessments. To this end, we created a sci-
ence achievement test containing NELS:88, TIMSS, and NAEP multiple-choice items (see Table 1 in Shavelson et al., this issue) and administered three performance assessments that, on the basis of a logical analysis of the task, response requirements, and scoring system, roughly fit one or another of the reasoning dimensions. In addition, we selected the 13 NELS:88 multiple-choice items on our test reflecting the three reasoning dimensions and examined the extent to which they fit the original three-factor reasoning structure found by Snow and colleagues. Finally, we examined the convergence of the multiple-choice items and performance assessments on the three reasoning dimensions. To be sure, performance assessments are far more complex than multiple-choice items (Li & Shavelson, 2001; Shavelson & Ruiz-Primo, 1999). Consequently, we recognized that they would tap other aspects of students’ knowledge and reasoning but nevertheless expected to find some combination of the three types of reasoning present in the performance assessment scores.

Performance Assessment Selection

To see whether these reasoning dimensions generalized to performance assessment scores, we reviewed and classified previously published assessments into one or another of the three reasoning dimensions (Ayala, Ayala, & Shavelson, 2000). We did this by examining their task demands, response demands, and scoring systems. For example, the “Paper Towels” performance assessment (Baxter, Shavelson, Goldman, & Pine, 1992) asked students to determine which of three paper towels absorbed the most—least water; performance was scored for the scientific justifiability of the procedures used in conducting the investigation as well as the accuracy of the students’ inferences based on the data they collected. We considered Paper Towels to tap basic knowledge and reasoning because it involved general science experimentation, general reasoning, and representations. Paper Towels, then, tapped both procedural knowledge (conducting a controlled investigation) and reasoning; however, no specific science content knowledge was required.

Twenty-seven performance assessments were analyzed by this method. Twenty-five assessments were classified as primarily basic knowledge and reasoning, 2 were classified as primarily spatial—mechanical reasoning, and none were classified as quantitative science (A. Ruiz-Primo, personal communication, September 21, 1999). To fill the quantitative science void, we developed a new performance assessment (Ayala et al., 2000).

As performance assessments are more complex than multiple-choice items, we did not expect a one-to-one correspondence between Snow’s reasoning dimensions and the performance assessments. Rather, we expected a preponderance of the reasoning underlying competent performance on a performance assessment to fit the reasoning dimension into which the assessment was placed. We speak, then,
of our classifications as "primarily" because, given the complexity of performance assessments, multiple types of reasoning and knowledge are most likely tapped. For example, all surely tap, to a greater or lesser degree, basic (procedural) knowledge and reasoning (e.g., Li & Shavelson, 2001).

In selecting performance assessments to represent the three reasoning dimensions, we also sought assessments that fell into the content-rich and process-open quadrant of Baxter and Glaser's (1998) Content-Process Space. This quadrant was expected to produce the most scientific reasoning. A performance assessment is content rich if it requires specific content knowledge to succeed. It is process-open if students formulate their own procedures for carrying out an investigation rather than following a procedure or "recipe."

Moreover, because reasoning demands are related to tasks (Baxter & Glaser, 1998), we characterized the assessments by the task types defined by Shavelson, Solano-Flores, and Ruiz-Primo (1998): (a) **comparative** investigations, in which students compare two or more objects and their performance is evaluated for accuracy of procedures and inferences; (b) **component-identification** investigations, in which the task is to decompose a whole (e.g., Electric Mystery box) into its components parts (wire, battery, and bulb, etc.) by connecting an external circuit, and performance is evaluated as to confirming and disconfirming evidence; (c) **taxonomic** investigations, in which students construct a taxonomy for a particular purpose such as predicting which objects would sink or float based on volume and mass, with performance evaluated as the accuracy and relevancy of the classifications; and (d) **observation** investigations, in which students observe and model a process over time, and their performance is evaluated as to the accuracy of the observations, models, and inferences. At a later date, we plan to compare reasoning dimensions and task types.

**Basic knowledge and reasoning dimension.** We selected Electric Mysteries as tapping primarily basic knowledge and reasoning. Students needed to know about the flow of electricity in a series circuit and had to reason about confirming and disconfirming evidence to perform the task successfully (Rosenquist, Shavelson, & Ruiz-Primo, 2000; Shavelson, Baxter, & Pine, 1991).

In this assessment, students were given batteries, bulbs, and wires and were asked to connect them to each of six "mystery" boxes to determine the boxes' contents—wire, nothing, two batteries, and so on (see Figure 1). Each task used the same equipment, and no procedures were given to students about how to complete the task, organize data, or find the correct solution. These six boxes were interchangeable, although some circuits were more difficult to identify than others. The scoring form asked raters to evaluate the circuit used to determine the contents of each box and to determine whether the student accurately determined the boxes' contents (all scoring forms are available at http://www.stanford.edu/dept/SUSE/SEAL/). Scoring
was straightforward and reliable (intrarater reliability was > .90; internal consistency was .81).

**Spatial–mechanical reasoning dimension.** We selected Daytime Astronomy as tapping primarily spatial–mechanical reasoning. Students were given an Earth globe in a box, a flashlight, and a set of “sticky towers” (see Figure 2). Students then used the flashlight as if it were the Sun to project shadows with the sticky towers to determine the time and location of places on Earth. Successful task performance required spatial observation, modeling, and reasoning (Solano-Flores, Jovanovic, & Shavelson, 1994; Solano-Flores & Shavelson, 1997; Solano-Flores et al., 1997), features of the spatial–mechanical reasoning dimension. Daytime Astronomy was considered content rich, requiring knowledge of the Sun’s position in relation to Earth, the Earth’s rotation, and the relation between the position of the Sun and shadows cast on Earth. Because students were not given directions on how to carry out the investigation, the assessment was considered process-open. Finally, as students were asked to model the path of the Sun across the sky and to use the direction, length, and angle of shadows, cast by sticky towers to solve location problems, Solano-Flores and Shavelson (1997) considered this assessment to be an observation task.

Daytime Astronomy was divided into six subtasks, some more closely related than others, but all designed to tap into a student’s understanding of the motion of the Sun in relation to the Earth and the shadows that this relation produces. Subtask 1 asked students to find the location of a sticky tower based on the length and orientation of its shadow in relation to the shadows of two other fixed towers. Subtask 2 asked students to determine shadow length, orientation, and movement for a tower
at 10 a.m. and 3 p.m. In Subtask 3, students determined the time of day and shadow length and orientation for a tower located in Seattle given that two towers located in the Midwest are at noon. Subtask 4 prompted students to explain the relation between Sun shadows and the time of day. Subtask 5 had students determine Sun shadow length, orientation, and movement for a tower at 10 a.m., noon, and 3 p.m. in the Southern Hemisphere. Finally, Subtask 6 asked students to describe the differences and similarities between Sun shadows in the Northern and Southern Hemispheres. Solano-Flores and Shavelson (1997) reported an interrater reliability of .90 on the scoring of Daytime Astronomy.

Quantitative science dimension. Finally, we developed a new investigation, Aquacraft, to tap, among other things, quantitative science. Students were asked to determine the cause of an explosion aboard a submarine by simulating what might have happened, when copper sulfate was added to aluminum ballast tanks using: glassware, copper sulfate, aluminum, salt, and matches (see Figure 3). To perform the task, students had to test for unknown gases, manipulate numerical quantities, and use specialized course-based knowledge—the general characteristics of the quantitative science dimension.

Students determined the cause of an explosion using high school chemistry principles and procedures, selected the appropriate chemical equations to represent the reaction, and determined quantitatively the amount of energy released in the explosion. Because advanced science content knowledge and specialized skills were needed to complete the task, it was considered content-rich. And because students conducted their own investigations without step-by-step instructions, it was considered process-open. Finally, because students were expected to
compared chemical reactions in both fresh and salt water, we considered Aquacraft to be a comparative investigation.

Aquacraft consisted of four subtasks. Subtask 1, *chemical reaction*, and Subtask 2, *test the gas*, consisted of observing and comparing two possible scenarios (copper sulfate in fresh water vs. copper sulfate in salt water) using chemistry procedures and lab techniques. Subtask 3, *balancing equations*, required students to select the appropriate chemical equations for the reaction. Subtask 4, *energy calculations*, asked students to determine quantitatively whether there was enough energy released in the explosion to cause the reported damage. All four subtasks tapped into the quantitative science reasoning dimension in different ways, the first and second via chemistry content and lab techniques, and the third and fourth via chemistry content and quantitative procedures. The scoring form evaluated students' procedures, observations, and conclusions, and in the last two tasks evaluated quantitative steps, explanations, and conclusions. Interrater reliability was .97.

The three performance assessments selected for our study and their classifications are summarized in Table 1. Once the assessments were selected, we examined their appropriateness for this study by administering them in a small pilot study (Ayala et al., 2000).

In this study, then, we ask (a) Do Snow and colleagues' finding of three reasoning dimensions underlying the NELS:88 multiple-choice science achievement test generalize to other multiple-choice tests? (b) Can these reasoning di-
TABLE 1
Performance Assessment Characteristics Based on the Three Frameworks

<table>
<thead>
<tr>
<th>Performance Assessment</th>
<th>Primary Reasoning Dimension</th>
<th>Content</th>
<th>Process</th>
<th>Task Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electric mysteries</td>
<td>Basic knowledge and reasoning</td>
<td>Rich</td>
<td>Open</td>
<td>Component-identification investigation</td>
</tr>
<tr>
<td>Daytime astronomy</td>
<td>Spatial-mechanical</td>
<td>Rich</td>
<td>Open</td>
<td>Observation investigation</td>
</tr>
<tr>
<td>Aquacraft</td>
<td>Quantitative science</td>
<td>Rich</td>
<td>Open</td>
<td>Comparative</td>
</tr>
</tbody>
</table>

Dimensions be found in other tests such as performance assessments, recognizing their complexity? and (c) Do the two different science achievement measurement methods converge on the three reasoning dimensions, lending support to the interpretation that these dimensions are implicated in science achievement?

METHODS

Respondents

We selected respondents in the following way. Of the 491 students participating in the main study, 343 had complete data on the achievement test; of those, 225 had complete data and indicated a willingness to participate in the performance assessment portion of the study. We recruited 35 of these students in the summer of 2000, following the main study. Because many of the 225 students were from the higher academic track, we focused our recruitment on lower performing students in this subsample. Care was also taken to select some students who had not yet completed high school chemistry and to select a sample that included girls. It was not easy to get students to take 2 hr of exams during the summer, and many students had to be reinvited before they came. Jokingly, one student remarked, “the hardest test was coming here.” But with all that, students from a variety of grade levels and achievement levels completed the performance assessments (Table 2).

We recognize that the sample size is small and not necessarily representative of the original sample in the main study in that their achievement was slightly higher than that of their peers (see later). However, the cost of data collection and scoring were limiting factors, as was availability of students. This study,
TABLE 2
Characteristics of Performance Assessment Respondents

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre chemistry</td>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Post chemistry</td>
<td>12</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>20</td>
<td>35</td>
</tr>
</tbody>
</table>

then, is but a small first step in investigating the reasoning dimensions underlying alternative assessments and it follows in the tradition of Baxter and Glaser (1998) and Ruiz-Primo, Shavelson, Li, and Schultz (2001).

Instrumentation

Students in the full sample completed a series of motivational and cognitive tests and surveys as well as science achievement measures (see Shavelson et al., this issue). Our focus here is on reasoning underlying science achievement—the 30-item multiple-choice test that contained items from NELS:88 (13 items), NAEP (6 items), and TIMSS (11 items)—and the three performance assessments. We included NAEP and TIMSS items, following Snow, to see whether the three reasoning dimensions were unique to NELS:88 or could be found in other multiple-choice tests, such as the one we constructed for this study.

The mean score for the 343 students who completed the multiple-choice test was 16.17 out of 30 (SD = 5.65). This mean score was surprisingly low considering that most students had already completed 3 years of high school science. Compared to the full sample, our performance assessment subsample (n = 35) had a slightly higher mean score on their multiple-choice test, 18.35 (SD = 4.54), t = 3.08, p < .05. We concluded that the subsample was somewhat higher on science achievement than the students in the main study.

Multiple-Choice Test Structure and Reliability

Snow sought to determine whether the reasoning dimensions underlying science achievement on NELS:88 generalized to other achievement tests. For this reason, we selected 13 NELS:88 items that had been shown to load strongly on one or another of the three factors—basic knowledge and reasoning (BKR), spatial–mechanical reasoning (SM), and quantitative science (QS). We also selected items from TIMSS and NAEP that, based on a task analysis of each item, tapped primarily one or another factor (see Table 1 in Shavelson et al., this issue, for item descriptions and classifications).

We examined the structure of the 30-item multiple-choice test and the subset of 13 NELS:88 items with data from the full sample of 343 students in the main study.
using confirmatory factor analysis. We posited three correlated reasoning dimensions corresponding to Snow and colleagues’ (Nussbaum et al., 1997) finding with NELS:88 and fit the item data.

Performance Assessment Classification, Scoring, and Reliability

Check on assessment classification into reasoning dimensions. To check our classification of each of the three performance assessments into one of the three reasoning dimensions, we conducted a small, "think-aloud" pilot study with three teachers ("experts") and three students ("novices"). Each of the performance assessments was administered individually to one expert (science teacher) and one novice (high school physics student); cost and time prohibited larger numbers. We reasoned that, if the nature of the assessment task had an effect on reasoning, prior research on expertise (e.g., Chi, Glaser, & Farr, 1988) suggested that using this extreme group design would allow us to detect the effect. Although every person constructs a somewhat different problem space when confronted with the same task environment (cf. Newell & Simon, 1972), experts are consistent in their substantive representations of the principle underlying the task, whereas novices are strongly influenced by the surface features of the task. Of course, the next step in this research would be to confirm systematic effects, if found, with multiple experts and novices.

Expert volunteers were assigned to the performance assessment that most closely matched their teaching expertise. A female chemistry teacher, with 4 years of teaching experience, was assigned Aquacroft; a female physical science teacher, with 7 years of teaching experience, was assigned Electric Mysteries; and a male general science teacher, with 13 years of experience, was assigned Daytime Astronomy. Student volunteers were randomly assigned to each of the different tests. All students were male high school physics students who had completed at least 2 years of high school science. The student assigned to Electric Mysteries was the only student who had not completed the high school chemistry course.

Students and teachers were asked to concurrently think aloud while they completed each performance assessment. Think-alouds were audio taped and transcribed. Similar procedures have been used before to investigate cognitive task demands of assessments (Baxter & Glaser, 1998; Ericsson & Simon, 1993; Ruiz-Primo, 1999).

We segmented and coded the verbal protocols in a manner similar to studies by Ericsson and Simon (1993). The think-alouds were segmented, and iterations of the coding categories were tried out on the segments. We coded segments that dealt with science procedures and reasoning (e.g., conducting, observing, concluding, conjecturing, explaining, planning) to reasoning dimensions; these seg-
ments reflected students' science knowledge and thinking. We did not code segments to reasoning dimensions if they pertained to reading test directions, test-taking comments ("turning the page" or "writing down the answer"), or exclamations ("holy cow, Batman" or "this is great"). As part of training and coding system development, two raters classified random segments of the think-alouds independently. The raters then discussed disagreements in coding and ways to make the segments more identifiable, the coding categories more explicit, or both. Finally, comparisons were made between the types of reasoning elicited from the performance assessments across novices and experts.

Differences in reasoning demands were evident in the think-aloud data. As expected, we found averaging over experts and novices that all three assessments drew on basic knowledge and reasoning, less so for Aquacraft than for the other two assessments. More specifically, Electric Mysteries drew almost solely on basic knowledge and reasoning (no think-aloud segment was coded spatial–mechanical or quantitative science). However, as expected, we found evidence of spatial–mechanical reasoning with Daytime Astronomy even though a greater percentage of the think-aloud segments were classified basic knowledge and reasoning (27%) than spatial–mechanical (20%). As designed, Aquacraft tapped primarily quantitative science (20% of statements) but also basic knowledge and reasoning (12%). The remainder of the segments were not identified as science procedures and reasoning segments and were not coded to reasoning dimensions. These limited data, then, supported our nominal analysis that Electric Mysteries tapped basic knowledge and reasoning, Daytime Astronomy tapped spatial–mechanical reasoning, and Aquacraft tapped quantitative science reasoning.

We concluded that, given the complexity of performance assessments, there was some evidence to support our classification: The performance assessments varied nominally in their demands on basic knowledge and reasoning, quantitative science, and spatial–mechanical reasoning. By selecting performance assessments using the general characteristics of the reasoning dimensions, and then collecting think-aloud protocols to study the reasoning these tasks evoked, we found that the different performance assessments did elicit the reasoning we expected, with different reasoning patterns found for the experts-and-novices study. Consequently, we proceeded with the performance assessment study.

_Electric Mysteries._ Two trained raters scored each of the Electric Mysteries performance assessments. Raters evaluated students' performance from their drawings of the circuit they used to conclude the contents of each Electric Mystery box and whether or not the student correctly identified the boxes' contents. Two raters scored each of the Electric Mysteries performance assessments with complete agreement.
**Daytime Astronomy.** Two trained raters scored the Daytime Astronomy performance assessment. A random sample of 13 of the 35 student notebooks was used for training purposes. Once raters had scored 13 notebooks and had discussed problematic areas, the remaining 22 notebooks were scored independently (intrarater reliability = .90).

**Aquacraft.** Two trained raters scored each of the Aquacraft performance assessments. A random sample of 10 of the 35 performance assessments was used for training purposes. Once the raters had reviewed the performance assessments used for training and discussed problematic areas, the remaining 25 performance assessments were scored independently (intrarater reliability = .97).

**RESULTS AND DISCUSSION**

**Multiple-Choice Tests**

If Snow's dimensions were not an artifact of the NELS:88 test, we should expect to find three factors corresponding to these dimensions in our 30-item science achievement test. This is, indeed, what we found. The items fit the three-factor model reasonably well: \( \chi^2 = 420.803, df = 321, \) Goodness-of-Fit Index (GFI) = .925, Adjusted Goodness-of-Fit Index (AGFI) = .911. The reasoning dimensions, however, were highly correlated: QS–BKR = .96, QS–SM = .82, and BKR–SM = .92. The evidence, then, was that these reasoning dimensions might be found in a multiple-choice test other than NELS:88. But the structure of our 30-item test with these high factor correlations was not well enough distinguished for the purposes of studying the convergence of multiple-choice and performance tests on the reasoning dimensions.

As a consequence, we examined the structure of the 13 NELS:88 items alone and confirmed in our sample Snow and colleagues' prior findings of three factors, \( \chi^2 = 59.621, df = 62, \) GFI = .977, AGFI = .966. The correlations among the factors were QS–BKR = .83, QS–SM = .85, and BKR–SM = .85. Because the factor correlations were lower with the NELS:88 13-item achievement test, we decided to use it for the study of convergence. The reliability of the test was total score (13 items) = .77, BKR (5 items) = .49, SM (3 items) = .65, and QS (5 items) = .61. In the following, we report findings based on the NELS:88 science achievement items. Incidentally, the fit indexes indicate that the three-factor model was a reasonable representation of the dimensional structure of the test, taking into account all items. In other words, a three-factor model is better than a unidimensional depiction. The extent to which each dimension can be measured reliably (at the individual student level) is a different question. The reliability
analyses were done separately for each dimension, and those with fewer items (like SM) were naturally less reliable.

Performance Assessment Scores

**Electric Mysteries.** The Electric Mysteries mean score for students was 28.1 out of a maximum of 48 points (Table 3). We disaggregated the scores into the four scoring categories to locate the source of students' errors. The new scoring form adjusted for the lack of drawings and credited explanations; however, students did not provide good explanations in their notebooks. The low score was largely due to the low mean explanation score, 2.39 out of 12 possible points. The drawing, observation, and inference subscores were much higher than the explanation subscores.

**Daytime Astronomy.** The Daytime Astronomy mean total score was 27.1 \((SD = 8.75)\). We do not report a maximum score for Daytime Astronomy because the scoring form was designed to capture all possible maneuvers in the problem solution path. Consequently, the total possible score was much higher than any single solution path would yield. For comparison, Shavelson et al. (1998) found that the Daytime Astronomy mean score with fifth graders was 14, and Ayala et al.'s (2000) Daytime Astronomy high school novice scored 34, whereas the expert scored 60. We used 60, then, as a benchmark.

Because the Daytime Astronomy scoring form contained three scoring categories, the data were disaggregated (Table 4). These scores reflected the accuracy of the students' observations, the quality and type of modeling that the students used in answering the questions, and the quality of their explanations.

**Aquacraft.** The Aquacraft mean total score was 15.7 out of a maximum of 42 (Table 5). For comparison, in our previous study, the Aquacraft novice scored 17 and the Aquacraft expert scored 32 out of a possible 42 (Ayala et al., 2000). One student (an outlier) scored very high, 35. This was the only student to solve

| TABLE 3 | Electric Mysteries Total Score and Subscores |
|-----------------|---------------------------------|-----------------|-----------------|-----------------|
|                | **Total** \((Max = 48)\) | **Drawing** \((n_i = 6)\) | **Observation** \((n_i = 6)\) | **Inference** \((n_i = 6)\) | **Explanation** \((n_i = 6)\) |
| **M**          | 28.10                          | 8.55                                   | 7.77                                   | 9.49                               | 2.39                        |
| **SD**         | 10.25                          | 3.84                                   | 3.14                                   | 3.10                               | 2.29                        |
TABLE 4
Daytime Astronomy Total Score and Subscores

<table>
<thead>
<tr>
<th></th>
<th>Daytime Astronomy Subscores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Results (n_i = 6)</td>
</tr>
<tr>
<td>Total</td>
<td>Benchmark = 60</td>
</tr>
<tr>
<td>M</td>
<td>27.10</td>
</tr>
<tr>
<td>SD</td>
<td>8.76</td>
</tr>
</tbody>
</table>

TABLE 5
Aquacraft Total Score and Subscores

<table>
<thead>
<tr>
<th></th>
<th>Aquacraft Subscores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Task 1 Chemical Reaction (n_i = 12)</td>
</tr>
<tr>
<td>Total</td>
<td>Max = 42</td>
</tr>
<tr>
<td>M</td>
<td>15.70</td>
</tr>
<tr>
<td>SD</td>
<td>6.97</td>
</tr>
</tbody>
</table>

all the problems; yet he did not get a perfect score because he failed to make all
the observations to compare the salt- and fresh-water conditions.

Further analysis of Aquacraft revealed that students tended to do better on
Subtask 3 (balancing equations; 6.16 out of 8 maximum) than on Subtask 4 (energy calculations; 2.20 out of 10 maximum). Both of these tasks required stu-
dents to use multiple calculations using content knowledge from 1st-year chem-
istry. Students clearly knew the process of balancing equations, and many were
able to identify elements from memory (i.e., they identified Ba, naming it bar-
im). However, students were not able to do the energy calculations. On review-
ing their notebooks, it appeared that they were not able to convert kilograms to
moles; that is, kg \times 1000 \text{ g/kg} \times (1/\text{atomic weight}) \text{ moles/g} = \text{moles}. Whereas
reasoning dimensions might be useful for explaining the types of processes used
by students to solve problems, as expected, the declarative and procedural
knowledge that these students had also played into the mix, and in the case of
Subtask 4 (energy calculations), this lack of procedural (algorithmic) knowledge
was limiting.

Correlations Between Performance Assessment and Total
NELS:88 Scores

Although the three performance assessments tapped basic knowledge and rea-
soning and procedural knowledge, they also measured different reasoning di-
mensions, as well as different aspects of declarative (content) and schematic ("mental model") knowledge. Consequently, we expected and found moderate correlations between performance assessment scores (Table 6). We also expected moderate correlations with the NELS:88 total score, and this is what we found for Electric Mysteries. The correlations between the NELS:88 total scores and the other two performance assessment scores were higher than anticipated. The magnitude of these two correlations (.67 for Daytime Astronomy and .51 for Aquacraft) most likely reflected the complexity and breadth of knowledge tapped by the multiple-choice total score and by these two performance assessments.

Convergence of Multiple-Choice and Performance Assessment Scores on the Reasoning Dimensions

We asked the question, To what extent do the NELS:88 multiple-choice reasoning subscales and the performance assessments converge on the reasoning dimensions they were intended to tap? We recognized that, because of their complexity, performance assessments might very well tap more than one reasoning dimension as well as various types of knowledge. Consequently, we expected to find not only convergence due to the “preponderance” of reasoning underlying one or another of the performance assessments, but also some “scatter” as the assessments tapped multiple types of knowledge and reasoning. To this end, we correlated the multiple-choice and performance assessment scores and arrayed them in a multireasoning-multimethod correlation matrix (Table 7).

To interpret the pattern of correlations bearing on convergence, we note that the main diagonal in Table 7 represents the reliabilities (in parentheses) of the corresponding measures. They set the limits on the magnitude of the correlations in the remainder of the table.

<table>
<thead>
<tr>
<th></th>
<th>Electric Mysteries</th>
<th>Daytime Astronomy</th>
<th>Aquacraft</th>
<th>NELS:88 Total Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electric Mysteries</td>
<td>.81*</td>
<td>.26</td>
<td>.37</td>
<td>.38</td>
</tr>
<tr>
<td>Daytime Astronomy</td>
<td>.90*</td>
<td>.51*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aquacraft</td>
<td></td>
<td></td>
<td>(.97)*</td>
<td>.51*</td>
</tr>
</tbody>
</table>

*Generalizability coefficient.

*Simultaneous confidence interval $r > .45$ significant at .05.
**TABLE 7**  
Multitasking—Multimethod Correlation Matrix (N = 35)

<table>
<thead>
<tr>
<th></th>
<th>Multiple Choice</th>
<th>Performance Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basic Knowledge (BKR)</td>
<td>Spatial-Mechanical (SM)</td>
</tr>
<tr>
<td>Multiple choice</td>
<td>BKR (.49)*</td>
<td>.42 (.65)*</td>
</tr>
<tr>
<td>SM</td>
<td>.42 (.65)*</td>
<td>.39 (.61)*</td>
</tr>
<tr>
<td>QS</td>
<td>.35 (.61)*</td>
<td></td>
</tr>
<tr>
<td>Performance assessment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electric Mysteries (BKR)</td>
<td>.34 .27 .29 (.99)b</td>
<td>.26 (.90)b</td>
</tr>
<tr>
<td>Daytime Astronomy (SM)</td>
<td>.41 .34 .66* (.90)b</td>
<td>.26 (.90)b</td>
</tr>
<tr>
<td>Aquacraft (QS)</td>
<td>.34 .51* (.97)b</td>
<td>.51* (.97)b</td>
</tr>
</tbody>
</table>

*Note. Underline denotes convergent validity coefficients.*

*Internal consistency reliability (N = 371), n(BKR) = 5, n(QS) = 5, n(SM) = 3. *Interrater reliability coefficients (N = 35). *Using simultaneous confidence interval, r > .51 significant at .05.

We now focus on the convergent validity diagonal—the diagonal with underlined correlations. These are validity coefficients—correlations between reasoning scores (e.g., BKR)—measured by two different methods: multiple choice and performance assessment (r = .34). If Snow’s three reasoning dimensions generalize beyond the NELS:88 multiple-choice test, we should expect the three validity coefficients to be statistically significant and of practical value. However, due to small sample size and the number of correlations considered, the only validity coefficient that meets this criterion is that for quantitative science. As expected, Aquacraft and NELS:88 quantitative science scores converged (r = .51). Nevertheless, the other two validity coefficients were positive and moderate in magnitude. Most puzzling was the correlation between Daytime Astronomy and the NELS:88 measure of quantitative science (r = .66). Very few (2%) of the Aquacraft think-aloud segments in the pilot study were classified as quantitative science.

A second criterion for the validity diagonal is that its entries are greater in magnitude than the corresponding row and column entries. For example, we should expect the convergent validity coefficient for basic knowledge and reasoning (.34) to be greater than the entries in its corresponding row (.27 and .29) and in its corresponding column (.41 and .34). None of the validity coefficients survived this criterion.

The lack of convergence for basic knowledge and reasoning might be explained, in part, by the pilot study finding that Daytime Astronomy tapped basic
knowledge and reasoning as much as or more than it did spatial–mechanical reasoning. The lack of convergence for spatial–mechanical reasoning might also be explained, in part, by Daytime Astronomy’s demand on basic knowledge and reasoning. Moreover, from our observations of the students completing Daytime Astronomy and from reading their notebooks, we found that students were indeed involved in spatial–mechanical activities, shining the flashlight from above, observing the change of shadow length and angle as the globe was rotated. But we also found that students used other content knowledge and reasoning to perform the assessment as well, including time zones, geometry, knowledge of meridians, longitude and latitude, and personal travel experience. For example, one student used time zones to place the tower in the correct location: “I know that the time [difference] between the Midwest and Seattle is about two hours ... and so aim tower [shadow] to make sure it was two hours behind instead of ahead.” This makes the assignment of Daytime Astronomy to any combination of reasoning dimensions questionable because the reasoning used by the student depends on the content knowledge and life history he or she brings to the task.

We are at loggerheads to explain the quantitative science finding. From our observations of students completing Aquacraft and from their notebooks, we found that students consistently performed well on Subtask 3 (balancing equations) and poorly on Subtask 4 (energy calculations). It may be that some students’ experience and knowledge in Subtask 3 (balancing equations) moved the subtask from a quantitative science task into a basic knowledge and reasoning task. To check this reasoning, we correlated Aquacraft subtask scores with the NELS:88 reasoning scores. We found negligible correlations among Subtasks 1 and 2 and the NELS:88 reasoning scores, and correlations of .45 and .56 between Subtask 3 and NELS:88 BKR and QS scores, respectively. This pattern of correlations provides some support for our observations of students’ performance. For Subtask 4, some students lacked experience and knowledge in performing quantitative energy calculations, which influenced their reasoning; they may have been grasping at straws to answer Subtask 4 (energy calculations). For example, one student became bogged down in simple metric conversions: “200 kilograms is 2 grams, 20 grams? I don’t remember. Oh. I’m thinking it’s 2 grams. Let me think. No, it’s not. No, it’s 2000 grams. Yeah, that’s what it is. So 200 kilograms is 2000 grams, I think.” The correlations between Subtask 4 scores and NELS:88 BKR, QS, and SM scores were, respectively, .14, .47, and .43.

Unfortunately, the jury is still out as to whether Snow’s three reasoning dimensions can be tapped by measurement methods other than multiple choice. We interpret our findings here as providing some optimism as well as a reality check. The optimism arises from the pattern of most of the correlations bearing on convergent validity with a small sample of 35 students. The reality comes in
the recognition of the complexity of performance assessments and the broad spectrum of reasoning and knowledge that they appear to tap.

CONCLUSIONS

Snow and colleagues (Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995; Hamilton et al., 1997; Nussbaum et al., 1997) found three reasoning dimensions underlying the NELS:88 multiple-choice science achievement test: basic knowledge and reasoning (BKR), spatial-mechanical reasoning (SM), and quantitative science (QS). Snow wondered just how representative these dimensions were of other multiple-choice tests, and whether they could be found in more complex tests such as performance assessments. In this study, we focused on Snow's interest in the generality of the reasoning dimensions underlying science achievement. More specifically, would these reasoning dimensions be found not only in the NELS:88 multiple-choice test items but also in other multiple-choice tests (e.g., NAEP or TIMSS) and alternative assessments, such as performance assessments?

In confirmatory factor analysis, we found the three reasoning dimensions in our 30-item multiple-choice test that incorporated items from NELS:88, NAEP, and TIMSS. We interpreted this finding as providing initial support for the generality of the dimensions in other multiple-choice tests. However, the three reasoning factors were highly correlated (range = .82 to .96). Further work is needed to see whether a multiple-choice test could be constructed specifically to fit Snow's reasoning dimensions and to test whether the dimensions do generalize, perhaps combining logical, factor analytic, and cognitive analysis in test building (cf. Li & Shavelson, 2001).

As a consequence of our structural findings for the 30-item test, we examined the factor structure of the 13 NELS:88 items in the 30-item test to see whether we could replicate Snow and colleagues' findings with a small set of items. We found that the structure could be replicated well with factor correlations somewhat lower in range (.83 to .85). However, the subscales corresponding to the three factors had few items (3 to 5 per scale) and, consequently, moderately low reliabilities (from .49 to .65). We used this NELS:88 multiple-choice test to examine convergence with performance assessments on the reasoning dimensions because of the lower factor correlations and motivated by the need to separate out the factors in a convergent study.

We identified three performance assessments that tapped, predominantly, one or another of Snow's reasoning dimensions: Electric Mysteries (BKR), Daytime Astronomy (SM), and Aquacraft (QS). We say predominantly because performance assessments are more complex than multiple-choice tests and, consequently, measure multiple reasoning and knowledge (e.g., declarative, procedural, and schematic) dimensions. From the main study, we enlisted 35 students, roughly representative of the total sample, who agreed to participate in the summer and carry out the perfor-
mance assessments. We recognize that this is a very small subset of the full sample of students and is not likely to be representative of the U.S. high school population, as was NELS:88; and so our findings on convergence are, at best, tentative.

To examine the convergence of multiple-choice and performance assessment scores on one or another of the three reasoning dimensions, we created a multireasoning–multimethod correlation matrix. The correlation between two measures of one reasoning dimension (e.g., NELS:88 BKR and Electric Mysteries) should be high if the measures are convergent. Further, these “convergent validity coefficients” should be higher than other correlations between one of these measures and a measure of another reasoning dimension (e.g., NELS:88–BKR correlation with Electric Mysteries should be higher than NELS:88–BKR with Daytime Astronomy). Our findings provided some optimism as well as a reality check. The optimism arose from the pattern of most of the correlations bearing on convergent validity with a small sample of 35 students. However, we could not explain why Daytime Astronomy (a predominantly SM assessment) correlated higher with NELS:88 QS items than did Aquacraft (a predominantly QS assessment). Moreover, there was evidence of scatter, with performance assessments correlating higher with measures of other dimensions than the one they “predominantly” selected measured. Reality, then, came in the recognition of the complexity of performance assessments and the broad spectrum of reasoning and knowledge that they tap. What we can tentatively conclude is that performance assessments do indeed tap the reasoning dimensions to a greater or lesser extent, but they tap other kinds of knowledge and reasoning as well. Detailed studies of the cognitive processing (reasoning and knowledge) demands of performance assessments and multiple-choice items would seem an important follow-up to this study (see, for example, Ayala, Yin, & Shavelson, 2002; Baxter & Glaser, 1998; Hamilton et al., 1997; Li & Shavelson, 2001; Yin, Ayala, & Shavelson, 2002).

ACKNOWLEDGMENTS

Carlos Cuauhtémoc Ayala is now at California State University, Sonoma.

This research was funded in part by a grant from the National Science Foundation (REC-9628293 “Multidimensional Student Assessments for High School Mathematics and Science”) and in part by the National Center for Research on Evaluation, Standards, and Student Testing, under the Educational Research and Development Centers Program, PR/Award Number R305B90002, as administered by the Office of Educational Research and Improvement, United States Department of Education.

The findings and opinions expressed in this article do not reflect the positions or policies of the National Science Foundation, the National Institute on Student
Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the United States Department of Education.

An earlier version of this article was presented at the annual meeting of the American Educational Research Association, Seattle, Washington, April 2001.

REFERENCES


