# A FRAMEWORK FOR ANALYZING REASONING IN WRITTEN ASSESSMENTS

Sam O. Nagashima
CRESST/University of California, Los Angeles

Nathaniel J.S. Brown
Indiana University

Alice Fu
Stanford University

Mike Timms
WestEd

Mark Wilson
University of California, Berkeley

Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, March 24-28, 2008.

**Introduction**

Science educators have long stressed the importance of students' ability to recognize how the scientific community uses arguments to construct knowledge (Driver, Newton, & Osborne, 2000; Duschl, Ellenbogen, & Erduran, 1999; Jimenez-Aleixandre, Rodrıguez, & Duschl, 2000; Kelly & Takao, 2002). Until recently, studies of argumentation in science classrooms have relied primarily upon Toulmin's Argument Pattern (TAP) as an analytic tool (Toulmin, 1958). For example, Russell (1981) utilized the framework in the analysis of teachers' presentations and representation of science during classroom instruction. Jimenez-Aleixandre & Diaz de Bustamante (1997) used TAP to examine students' discourse during classroom investigations. Despite its frequent use, TAP presents a number of limitations. Difficulties in reliably identifying the parts of an argument that fit into the various components of Toulmin's model have forced many researchers to focus on only limited aspects of the total structure of an argument (Erduran et al., 2004). Given these challenges, some have modified Toulmin's framework (e.g., Zohar and Nemet, 2002; Kelly and Takao, 2002), unfortunately, the resulting framework still struggles to address important complex relationships within the structure of a students' argument (Sampson et al., 2006).

Mindful of earlier work in students' use of evidence, this paper describes the application of the Using Evidence Framework (Brown et al., 2008; see Figure 1) in the development of written assessment items. The Using Evidence Framework contains two distinct classes of information, Component and Process. Component refers to statements that situate and frame the context of the argument and includes five pieces - Premise, Claim, Rule, Evidence, and Data.

The second types of information in the framework are the Process piece, which is composed of three parts - Application, Interpretation, and Analysis.

This study involved the creation of assessment items targeting the buoyancy comprehension of middle and high school students. Given that many of the current assessment systems are based on outdated theories about student learning (Pellegrino, Chudowsky, & Glaser; 2001) and fail to meet the challenges in systematically measuring complex reasoning in science, new items were constructed that explicitly elicit student use of evidence. These new items correspond with pieces of the Using Evidence Framework drawing on both the component and process pieces, and in doing so, provides opportunities to bring about the multiple elements involved in students' reasoning.

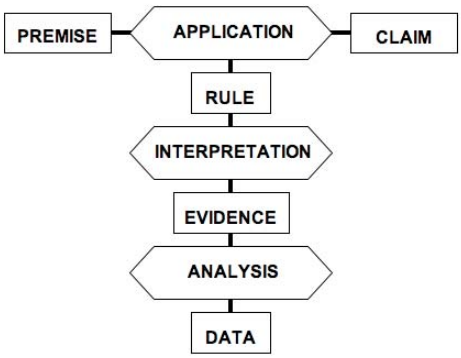



*Figure 1. The Using Evidence Framework*

**Methods**

In this study, we made use of the construct-mapping approach to items design (Wilson, 2005). The construct-mapping approach consists of four building blocks: a Construct Map, an Items Design, an Outcome Space, and a Measurement Model. These four building blocks form a

coherent system of assessment in which observations of student performance are interpreted with respect to a model of student cognition (Pellegrino, Chudowsky, & Glaser; 2001).

### *Construct Map*

Construct maps represent the particular concepts and skills that form the core learning goals of a curriculum, organized according to a developmental perspective of student learning in which deeper understandings are developed from, and take the place of, earlier understandings as students progress toward higher levels of sophistication and competence. In developing our construct maps, we focused on four aspects of using evidence that coincide with pieces of the framework: (1) *Conceptual Sophistication* of the Rule used by the student; (2) *Precision* of the Rule used by the student; (3) *Validity* of their Application of that Rule; (4) *Reliability* of the Evidence supporting that Rule.

The Conceptual Sophistication construct captures the quality and complexity of the concepts that the Rule implicates. Ranging from misconceptions, at the lowest level, to normative scientific conceptions that require the coordination of several ideas, Conceptual Sophistication addresses, directly, the complexity with which students represent concepts (Table 1). Conceptual Sophistication is an attempt at capturing the relationship between content matter and students' use of evidence.

Table 1. *Construct Map for Conceptual Sophistication.*

| *Levels* | | *Description* |
|---|---|---|
| 6 | Multi-Relational | Student use more than one normative, combined concept. |
| 5 | Combined | Student uses one normative concept that is a function of other concepts. |
| 4 | Relational | Student uses more than one normative concept. |
| 3 | Singular | Student uses one normative concept. |
| 2 | Productive Misconception | Student uses concepts that are not normative, but that can provide starting points for instruction. |
| 1 | Unproductive Misconception | Student uses concepts that are too far removed from normative to provide starting points for instruction. |

The Precision construct deals with the degree of specificity in the formulation of the Rule. Ranging from ambiguous values to explicit quantities with appropriate units, Precision was selected as a domain-general construct that is independent of content and used as a method for capturing the degree to exactness in the use of Rule (Table 2).

Table 2. *Construct Map for Precision.*

| *Level* | | *Description* |
|---|---|---|
| 4 | Exact | Response refers to a measurable or testable property in a way that defines a precise measurement or magnitude. |
| 3 | Inexact | Response refers to a measurable or testable property in a way that defines a measurement or magnitude but with a slight lack of precision. |
| 2 | Vague | Response refers to a measurable or testable property in a way that loosely defines its measurement or magnitude. |
| 1 | Indeterminate | Response refers to a measurable or testable property in a way that fails to define its measurement or magnitude. |

The third construct, Validity, examines the quality of the reasoning linking the Rule to the Claim. Ranging from no link to valid logical conclusions, Validity, like Precision, is domain independent and is an attempt to follow the line of reasoning and evaluate its legitimacy (Table 3).

Table 3. *Construct Map for Validity.*

| | *Level* | *Description* |
|---|---|---|
| 4 | Fully Valid | The outcome is consistent with the rule. |
| 3 | Partially Valid | The outcome is partially consistent with the rule. |
| 2 | Invalid | The outcome is inconsistent with the rule. |
| 1 | No Link | Response is impossible to categorize as valid or invalid because no in made with the rule. |

Lastly, the Reliability construct refers to the quality of the source and the quantity of the Data that makes up the Evidence, ranging from made-up examples to controlled experiments with multiple trials (Table 4). Collectively, these four construct maps cover content specific (Conceptual Sophistication) and content independent (Conceptual Precision, Validity, Reliability) dimensions. The selections of these constructs were a deliberate attempt to address gaps in earlier work. With the establishment of the dimensions of measure, we will now discuss the method used to inform the construction of items.

Table 4. *Construct map for Reliability.*

| | *Level* | *Description* |
|---|---|---|
| 3 | Understanding | Response shows a strong understanding of how reliability is achieved and its importance in scientific experiments/investigations. |
| 2 | Some Understanding | Response shows some understanding of how reliability is achieved and its importance in science experiments/investigations. |
| 1 | Limited Understanding | Response shows some understanding of how reliability is achieved and its importance in science experiments/investigations. |

*Items Design*

The Items Design is a framework for designing tasks to elicit specific kinds of evidence about student knowledge that can be interpreted in terms of the construct maps. Informed by existing assessment items on the topic of sinking and floating, our goal was to develop new items that tap into students' use of evidence and correspond to elements of the framework. However, early in the item development process, it became clear that an item linked to single element of the framework (e.g., Rule, Evidence) fails in capturing the complex quality of students' use of



Formatted: Keep with next
Formatted: Keep with next
Formatted: Keep with next
Formatted: Keep with next
Formatted: Keep with next
Formatted: Keep with next
Formatted: Keep with next
Formatted: Keep with next
Formatted: Keep with next

evidence. For example, an attempt to evaluate a student generated Rule in isolation of the Interpretation falls short in capturing the underlying rational for why a particular Rule was selected. It was concluded that simultaneous appraisal of multiple pieces of the framework is required to accurately assess a line of reasoning. Acknowledging this dependency across multiple pieces of the framework, items were designed using combinations of seven different process-component pairs (Table 5).

Working up from the bottom of the Using Evidence Framework (see Figure 1), the first process-component pair is Analysis-Evidence. Typically, these items provide Data and require students to perform Analysis to either produce Evidence or critique Evidence provided. For example, a student may be presented with data from experiments on sinking and floating from multiple sources and then asked to evaluate a statement about objects that sink (see item 2f in Appendix C). The next process-component pair is Interpretation-Rule. These items are designed to allow students, given Evidence, an opportunity to critique a Rule that is either provided or created (see item 5b in Appendix C). The third process-component pair is Application-Claim. Application-Claim items provide a Rule and a Premise requiring students to perform Application to either produce or critique a Claim (see item 4c in Appendix C). Lastly, Application-Premise items provide a rule and a Claim and require students to perform Application to either produce a Premise or critic a Premise that is provided.

Table 5. *Process-Component pairs used in designing items.*

| *Process-Component Pair* | *Description* |
|---|---|
| *Moving up the framework* | |
| Analysis-Evidence | These items provide Data and require the student to perform Analysis to either produce Evidence or critique Evidence that is provided. |
| Interpretation-Rule | These items provide Evidence and require the student to perform Interpretation to either produce a Rule or critique a Rule that is provided. |
| Application-Claim | These items provide a Rule and a Premise and require the student to perform Application to either produce a Claim or critique a Claim that is provided. |
| Application-Premise | These items provide a Rule and a Claim and require the student to perform Application to either produce a Premise or critique a Premise that is provided. |
| *Moving down the framework* | |
| Application-Rule | These items provide a Premise and a Claim and require the student to perform Application to either produce a Rule or critique a Rule that is provided. |
| Interpretation-Evidence | These items provide a Rule and require the student to perform Interpretation to either produce Evidence or critique Evidence that is provided. |
| Analysis-Data | These items provide Evidence and require the student to perform Analysis to either produce Data or critique Data that is provided. |

Moving down the Using Evidence Framework, a different set of process-component pairs emerge. Application-Rule items provide a Premise and a Claim and require the student to perform Application to either produce a Rule or Critique a Rule that is provided. For example, a student presented with a series of objects floating in a medium is asked to explain the Rule that governs the observed phenomena (see item 3a in Appendix C). Interpretation-Evidence Items provide a Rule and require the student to perform Interpretation to either produce Evidence or Critique Evidence that is provided for that Rule (see item 3c Appendix C). Finally, Analysis-

Data items provide Evidence and require the student to perform Analysis to either produce data or critique data.

Here we provide an example of the instrument at work. Question 3, shown in Table 6, contains three items that are aligned with two process-component pairs, Application-Rule and Interpretation-Evidence. Item 3a is scored using two constructs, Conceptual Sophistication and Precision, item 3b is scored on Validity and Reliability, and item 3c is scored only on Validity.

Table 6. *Sample Item.*

| *Item* | *Question* | *Process-Component Pair* | *Construct* |
|---|---|---|---|
| | Here are some things that float in water: a kitchen sponge, a plastic toy boat, an empty glass bottle | | |
| 3a | What do these things have in common that causes them to float in water? | Application-Rule | Conceptual Sophistication<br>Precision |
| 3b | Scientists require evidence to support their beliefs. Describe a specific thing you've seen, heard, or done that supports your belief that things float because of the reason you described in 3a. | Interpretation-Evidence | Validity<br>Reliability |
| 3c | Describe a specific thing, either real or imaginary, that would disprove your belief that things float because of the reason you described in 3a. | Interpretation-Evidence | Validity |

In a later section we will discuss a sample student response to this question, and show how the response can be coded using the relevant outcome space.

### *Outcome Space*

The outcome space describes the qualitatively different kinds of student response elicited by the items and maps these classes of response to the levels of the construct map. Outcome spaces were developed for each of the four construct maps described above: six response

categories for Conceptual Sophistication, four for both Precision and Validity, and three for Reliability.

For the Conceptual Sophistication outcome space, the concept of sinking and floating was unpacked into levels of increasing sophistication and mapped back to the original construct map (Table 7). The highest level of conceptual sophistication are exhibited in answers that express sinking and floating in terms of the density of the object in relation to the density of the medium. The lowest level of sophistication are responses that refer to attributes such as shapes and the presence of holes when describing properties related to sinking and floating. Similarly, outcome spaces for Precision (Table 8), Validity (Table 9), and Reliability (Table 10) were more clearly delineated and mapped back to their respective construct maps. Given that these last three constructs are domain independent, generally, the levels of performance tend to be broadly defined.

Table 7. *Outcome space for Conceptual Sophistication.*

| *Level* | *Sub-codes* | *Description* |
|---|---|---|
| Multi-Relational | DD: Density of object & liquid | Students know that floating depends on the object having less density than the medium. |
| Combined | DE: Density | Students know that floating depends on the object having a small density. |
| Relational | MV: Mass & Volume<br>OS: Omitted Subsurface<br>BG: Buoyancy and Gravity | Students know that floating depends on having a small mass and a large volume |
| Singular | VO: Volume<br>MA: Mass<br>BF: Buoyant Force | Students know that floating depends on having a small mass or small volume. |
| Productive Misconception | BR: Backwards Rule<br>WE: Weight or heavy, light, etc.<br>SZ: Size or big, small, etc.<br>AH: Air or Hollow, etc.<br>SA: Surface Area or area<br>BY: Buoyancy<br>PL: Productive Logic<br>OP: Other Productive<br>NC: No Concept | Students think that floating depends on having a small size, heft, or amount, or that it depends on being made out of a particular material. |
| Unproductive Misconception | SH: Shape<br>HO: Holes<br>UL: Unproductive logic<br>OU: Other Unproductive | Students think that floating depends on the shape of an object or the absence of holes. |

Table 8. *Outcome space for Precision.*

| *Level* | *Sub-codes* | *Description* |
|---|---|---|
| Exact | EX: Exact | Response refers to a measurable or testable property in a way that defines a precise measurement or magnitude. |
| Inexact | UN: Unit less or Wrong units<br>RV: Relative | Response refers to a measurable or testable property in a way that defines a measurement or magnitude but with a slight lack of precision. |
| Vague | VA: Vague | Response refers to a measurable or testable property in a way that loosely defines its measurement or magnitude. |
| Indeterminate | ME: Measurable<br>NM: NonMeasurable<br>UK: Unknown | Response refers to a measurable or testable property in a way that fails to define its measurement or magnitude. |

Table 9. *Outcome space for Validity.*

| *Level* | *Sub-codes* | *Description* |
|---|---|---|
| Fully Valid | FV: Fully valid | The outcome is consistent with the rule |
| Partially Valid | PV: Partially Valid | The outcome is partially consistent with the rule |
| Invalid | IV: Invalid | The outcome is inconsistent with the rule |
| No Link | NL: No Link<br>MU: Misunderstand | Response is impossible to categorize as valid or invalid because the magnitude of a property isn't specified |

Table 10. *Outcome space for Reliability.*

| *Level* | *Sub-code* | *Description* |
|---|---|---|
| Understanding | UR: Understanding | Student understands the nature of standards of reliability and can choose an appropriate standard |
| Some Understanding | SU: Some Understanding | Student uses enough evidence to meet a particular standard of reliability |
| Limited Understanding | LU: Limited Understanding | |

In addition to the primary response categories for each outcome space, a common set of low-level codes were used across all four dimensions to capture responses that were unscorable and/or did not fit into the levels (e.g., simply repeating the question, stating "I don't know") (Table 11).

Table 11. *Example of common low-level codes used across all four constructs.*

| Low-level | Description |
|---|---|
| IF: Indefeasible | Student provides an unchallengeable explanation. |
| PR: Prediction only | Student only provides a prediction with no explanation. |
| TA: Tautology | Student uses repetition of question as response |
| IK: I Know | Student simple state "I know" |
| PA: Pattern only | Student describes a pattern with no explanation. |
| DU: Didn't Understand | Student misunderstood the question. |
| BL: Blank | Answer left blank. |

With the item design and outcome space prepared, thirty-one items were developed using the component-process pairs as a guide. The final items address four pieces of the Use of Evidence Framework: two Component pieces, Evidence and Rule; and two Process pieces, Interpretation and Application. Although a concerted effort was made to evenly distribute items across all process-component pairs, not all process-component pairs were amenable to written assessment items. As a result, there is an imbalance in the number of items associated with each process-component pair (Table 12). Furthermore, some items draw on more than one process-component pair. For example, a student may be asked to produce a Rule, and in the process must evaluate evidence, in such a case, the items require Interpretation-Rule as well as Analysis-Evidence.

Table 12. *Number of items developed for each process-component pair.*

| *Process-Component Pair* | *Number of Items* |
|---|---|
| *Moving up the framework* | |
| Analysis-Evidence | 10 |
| Interpretation-Rule | 5 |
| Application-Claim | 6 |
| Application-Premise | 0 |
| *Moving down the framework* | |
| Application-Rule | 7 |
| Interpretation-Evidence | 2 |
| Analysis-Data | 0 |

Draft items were subject to think-aloud analysis in which a few students verbalized their thinking process as they solved the problems while a researcher captured their comments. Based

on feedback from the think-alouds, items were revised and compiled into two test forms (A & B) with overlapping items. The two test forms were then pilot-tested in four middle schools and one high school in the San Francisco Bay Area. A total of eight teachers, 17 classes, and 343 students participated; participating classes had studied buoyancy at some time during that academic year (see Appendix A). Six scorers were trained in the use of the outcome spaces and scored the student responses from the pilot test. All items were double-scored and, in some cases, triple-scored.

Many items elicit information on more than one dimension (e.g., if a student was presented with multiple pieces of Evidence and prompted to generate a Rule consistent with that Evidence, their Rule could be judged in terms of both Conceptual Sophistication and Precision). See appendix B for a summary of study sample and break down of items by dimension, rater, and form.

A sample student response to the example question 3 is provided on Table 13. Looking first at item 3a for Conceptual Sophistication, we can see that the student response associates the density of the object to the density of water. The student produces the Rule that objects with less density than water will float, so in this case, the item is scored at “Multi-relational” with sub-code “Density of object & liquid”.

Table 13. *Sample student response to sample item.*

| *Item* | *Student Response* |
|---|---|
| 3a | "All of these items are less dense than water which causes them to float" |
| 3b | "Waters density is 1.00g/ml if anythings density is less than that, it will float. We put objects who's density is less than water floated, but objects with a higher density will sink." |
| 3c | "If an items density was higher than waters and floated, that will disprove my belief. Also, is an items density was less than waters and sank, that will also disprove my belief." |

Looking at item 3a on the construct for Precision, we now critique the exactness of the Rule that the student brings to bear on his response. Again, the student response describes objects floating and sinking in terms of the relative density of the object to the density of water. Because the student establishes the Rule in relative terms, the student is scored at the level of "Inexact", with the sub-code "Relative" for the precision of the statement.

Moving onto item 3b, an item that addresses Interpretation-Evidence. The student is provided with a Rule, in this case his own response to item 3a, and is asked to produce evidence. In looking at this item on the Validity construct, the student response presents information that is in agreement with the Rule developed in item 3a, that is, all parts of the response are consistent with the rule, as a result, this response is scored as "Fully valid". Looking at the response using the Reliability construct, we must evaluate the reliability of the evidence that the student produces in support of the rule. The student response recalls an investigation with water and objects of known density, accordingly, this response is scored "Understanding".

The last item, item 3c, requires the student to bring to bear counter-evidence that, if valid, would undermine the response to 3a. The student states an opposite outcome to the experiment referenced in 3b, were objects with greater density water floated and object with less density sank. This item is scored as “Fully valid” given the consistency between the response to 3c and Rule in 3a.

*Measurement Model*

The measurement model provides estimates of person proficiency, item difficulty, and rater harshness calibrated onto interval scales using a multidimensional, multi-facet item response model (ConQuest: Adams, Wilson, & Wang, 1997). In this initial analysis, items were examined using a partial-credit model with the four dimensions of the construct (Conceptual Sophistication, Precision, Validity, and Reliability). The highest code of the multiple rating was used and all low level-codes were treated as missing data. A single main effect was estimated to take into account rater effects.

**Results**

The four dimensions showed moderate correlations in the range .52 to .75, providing evidence that these four aspects of the Framework can be meaningfully separated. Rater harshness parameters showed slight and non-statistically-significant differences, providing evidence that the outcome spaces were applied consistently across all six raters. Initial measures of person separation reliability were borderline unacceptable (0.7 to 0.8); this was attributed in part to three poorly performing items that have been subsequently dropped. Work is being conducted to establish the validity and reliability of the remaining items.

**Discussion**

Emerging results suggest that the Using Evidence Framework is useful in designing science assessment items that probe how students are considering and using evidence, an important aspect of science inquiry skills. This study showed that meaningful interpretations of several dimensions are possible, and that trained raters can be relied upon to accurately and consistently apply the tool to written classroom assessments. An important feature of utilizing written assessment in the evaluation students' use of evidence is the ability to focus on, and reveal, aspects of the structure of an argument that are of interest. In linking the assessment items to the framework, deliberate care has been to taken to ensure that both content specific and content neutral aspects of students' use of evidence are addressed. We believe this is an important step in better understanding how to purposefully craft highly targeted tools to capture features of students' use of evidence that could be missed in assessments otherwise.

**References**


Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. Applied Psychological Measurement, 21, 1-23.

Brown, N. J. S., Wilson, M. R., Nagashima, S. O., Timms, M., Schneider, S. A., & Herman, J. L. (2008). *A Model of Scientific Reasoning*. Paper presented at the American Educational Research Association.

Clark, D. B. and Sampson, V. D. 2005. Analyzing the quality of argumentation supported by personally-seeded discussions. In Proceedings of The 2005 Conference on Computer Support For Collaborative Learning: Learning 2005: the Next 10 Years! (Taipei, Taiwan, May 30 - June 04, 2005). Computer Support for Collaborative Learning. International Society of the Learning Sciences, 76-85.

Duschl, R., Ellenbogen, K., & Erduran, S. (1999, April). Understanding dialogic argumentation. Paper presented at the annual meeting of American Educational Research Association, Montreal.

Driver, R., Newton, P. and Osborne, J. (2000) Establishing the norms of scientific argumentation in classrooms. Science Education, 84(3), 287-312.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. Science Education, 88, 915--933.

Jimenez-Aleixandre, M., & Diaz de Bustamante, J. (1997). Analyzing classroom discourse: Practical work in the biology laboratory. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Jiménez-Aleixandre, M. P., Rodríguez, A. B., & Duschl, R. A. (2000). “Doing the lesson” or “doing science”: argument in high school genetics. Science Education, 84, 757-792.

Kelly, G. J., & Takao, A. (2002). Epistemic levels in argument: An analysis of university oceanography students’ use of evidence in writing. Science Education, 86, 314-342.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

Russell, T.L. (1983). Analyzing arguments in science classroom discourse: Can teachers' questions distort scientific authority? Journal of Research in Science Teaching, 20, 27-45.

Toulmin, S.E. (1958). The Uses of Argument. Cambridge: Cambridge University Press.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Erlbaum.

Zohar, A. & Nemet, F. (2002) Fostering students' knowledge and argumentation skills through dilemmas in human genetics. Journal of Research in Science Teaching, 39, 35-62.

## Appendix A: Summary of Study Sample

| | Location | Teacher | Class | Form A Students | Form B Students | Total Students |
|---|---|---|---|---|---|---|
| | Palo Alto | A | 1 | 11 | 10 | 21 |
| | | B | 2 | 12 | 10 | 22 |
| | | | 3 | 8 | 8 | 16 |
| | | | 4 | 9 | 8 | 17 |
| | | | 5 | 10 | 11 | 21 |
| | | | 6 | 10 | 9 | 19 |
| | East Palo Alto | C | 7 | 12 | 11 | 23 |
| | | D | 8 | 13 | 13 | 26 |
| | | | 9 | 14 | 14 | 28 |
| | Tracy | E | 10 | 12 | 10 | 22 |
| | | F | 11 | 13 | 11 | 24 |
| | Union City | G | 12 | 10 | 11 | 21 |
| | | | 13 | 8 | 4 | 12 |
| | | | 14 | 1 | 2 | 3 |
| | | | 15 | 8 | 11 | 19 |
| | | | 16 | 9 | 8 | 17 |
| | Atherton | H | 17 | 16 | 16 | 32 |
| Total # | 5 | 8 | 17 | 176 | 167 | 343 |

## Appendix B: Summary of Items

| | **Form A**<br>(Raters: A, B, C)<br>175 students | **Form B**<br>(Raters: D, E, F)<br>168 students |
|---|---|---|
| **Conceptual Sophistication** | 1ab (common to both forms)[2]<br>3a (common to both forms)[2]<br>4b (common to both forms)[1] | |
| | 4c[1]<br>5a[1]<br>6a[1]<br>6c[1]<br>7ab[2] | 4c[1]<br>6b[1] |
| **Precision** | 3a (common to both forms)[2]<br>4b (common to both forms)[1] | |
| | 4c[1]<br>5a[1]<br>6a[1]<br>6c[1]<br>7ab[2] | 4c[1]<br>6b[1]<br>6d[1]<br>6f[1] |
| **Validity** | 1ab (common to both forms)[2]<br>3b (common to both forms)[2]<br>3c (common to both forms)[2]<br>2f (common to both forms)[1]<br>4b (common to both forms)[1] | |
| | 4c[1]<br>5a[1]<br>5b[1]<br>6b[1]<br>6d[1]<br>7ab[2] | 4c[1]<br>6b[1]<br>6d[1]<br>6f[1] |
| **Reliability** | 2d (common to both forms)[1]<br>2e (common to both forms)[1]<br>3b (common to both forms)[2] | |

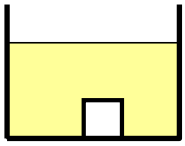
[1]For these items, raters stayed strictly within their Form A/B assignments: Form A raters always scored Form A responses; and Form B raters always scored Form B responses. All six raters participated in scoring.
[2]For these items, raters did *not* stay within their Form A/B assignments: some Form A raters scored Form B responses; and some Form B raters scored Form A responses. Only four raters participated in scoring.

| | **# Common Items Scored** | **# Total Items Scored** |
|---|---|---|
| Conceptual Sophistication | 3 | 10 |
| Precision | 2 | 11 |
| Validity | 5 | 15 |
| Reliability | 3 | 3 |

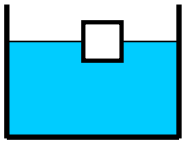
## Appendix C: Assessment Form A

Use the following information to answer Questions 1a and 1b.

A cube sinks in Liquid X.

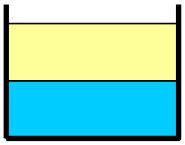



The same cube floats in water.




**1a. Which of the following must be true? Circle A or B.**

***[APPLICATION-PREMISE]***




A. Liquid X floats on water.


B. Liquid X sinks in water.

1b. Explain why you chose A or B.

***[APPLICATION-RULE, CS/V]***

Use the following information to answer Questions 2a, 2b, 2c, 2d, 2e, and 2f.

A teacher writes on the chalkboard, “Things sink the same amount whether they are in freshwater or saltwater.” She asks students to provide data for or against this statement.

Student A: In last week’s demonstration, the teacher put an egg in water, and it sunk. When the teacher added a lot of salt to the water, the egg floated.

**2a. Do the data from Student A support the teacher’s statement? Check YES or NO.**
***[LOW-LEVEL ANALYSIS-EVIDENCE]***

YES _____ NO _____

Student B: When I go swimming in the ocean in the summer, it doesn’t feel any different than when I go swimming in the pool. I think I sink the same amount whether I’m in the ocean or in the pool.

**2b. Do the data from Student B support the teacher’s statement? Check YES or NO.**
***[LOW-LEVEL ANALYSIS-EVIDENCE]***

YES _____ NO _____

Student C: Here is the data table from our experiment yesterday on four different objects.

| Object | Depth of sinking in water with no salt | Depth of sinking in water with some salt | Depth of sinking in water with a lot of salt |
|---|---|---|---|
| A | 8cm | 6cm | 4cm |
| B | 10cm | 7cm | 4cm |
| C | 5cm | 3cm | 1cm |
| D | 10cm | 9cm | 8cm |

**2c. Do the data from Student C support the teacher’s statement? Check YES or NO.**
***[LOW-LEVEL ANALYSIS-EVIDENCE]***

YES _____ NO _____

2d. Which of the students' data do you trust the most? Why?

*[ANALYSIS-EVIDENCE, R]*

2e. Which of the students' data do you trust the least? Why?

*[ANALYSIS-EVIDENCE, R]*

2f. Based on the data provided by all three students, do you think the teacher's statement is true? Why or why not?

*[ANALYSIS-EVIDENCE, INTERPRETATION-RULE, V]*

Use the following information to answer Questions 3a, 3b, and 3c.

Here are some things that float in water:

A. A kitchen sponge
B. A plastic toy boat
C. An empty glass bottle

3a. What do these things have in common that causes them to float in water?

***[APPLICATION-RULE, CS/P]***

**3b. Scientists require evidence to support their beliefs. Describe a specific thing you've seen, heard, or done that supports your belief that things float because of the reason you described in 3a.**

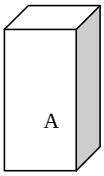
***[INTERPRETATION-EVIDENCE, V/R]***

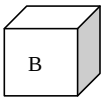
**3c. Describe a specific thing, either real or imaginary, that would disprove your belief that things float because of the reason you described in 3a.**

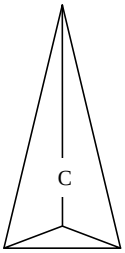
***[INTERPRETATION-EVIDENCE, V]***

Use the following information to answer Questions 4a, 4b, and 4c.

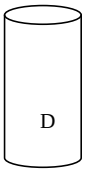
Pam places six objects in an unknown solution and records the results below.

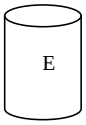


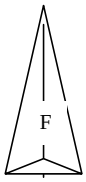




| | A | B | C |
|---|---|---|---|
| Mass: | 30 g | 20 g | 12 g |
| Volume: | 20 $cm^3$ | 10 $cm^3$ | 15 $cm^3$ |
| Density: | 1.5 $g/cm^3$ | 2.0 $g/cm^3$ | 0.8 $g/cm^3$ |
| | **Sink** | **Sink** | Float |




| | D | E | F |
|---|---|---|---|
| Mass: | 18 g | 32 g | 5 g |
| Volume: | 15 $cm^3$ | 8 $cm^3$ | 10 $cm^3$ |
| Density: | 1.2 $g/cm^3$ | 4.0 $g/cm^3$ | 0.5 $g/cm^3$ |
| | Float | **Sink** | Float |

**4a. Write 1-2 sentences about the patterns that you see in the data presented above.**

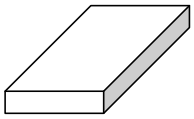
***[ANALYSIS-EVIDENCE, NOT SCORED]***

4b. Using your answer to 4a, state a general rule for what objects will sink in the unknown solution and what objects will float in the unknown solution.

*[INTERPRETATION-RULE, CS/P/V]*

**4c. Based on the rule that you provided in 4b, predict what will happen to the piece below when placed in the unknown solution. Will this piece sink or float? Why?**

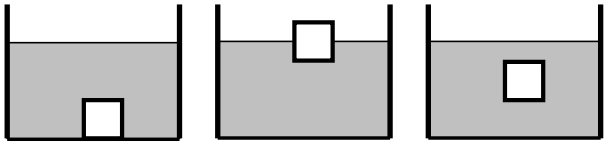
***[APPLICATION-CLAIM, APPLICATION-RULE, CS/P/V]***



Mass: 30 g
Volume: 10 $cm^3$
Density: 3.0 $g/cm^3$

Use the following information to answer Questions 5a and 5b.

A student investigates the effects of mass and volume on the sinking and floating of objects in water. She places several different objects in water to see if they will sink, float, or subsurface float.



Sink Float Subsurface Float

The data are shown in the following table.

| Object | Mass of Object | Volume of Object | Sink, Float, or Subsurface Float? |
|---|---|---|---|
| A | 8 g | 5 $cm^3$ | Sink |
| B | 15 g | 10 $cm^3$ | Sink |
| C | 15 g | 15 $cm^3$ | Subsurface Float |
| D | 12 g | 10 $cm^3$ | Sink |
| E | 20 g | 20 $cm^3$ | Subsurface Float |

**5a. Write 1-2 sentences about the patterns that you see in the data table.**

***[ANALYSIS-EVIDENCE, CS/P/V]***

Hypothesis 1: When the mass of an object is greater than its volume, the object will sink in water.

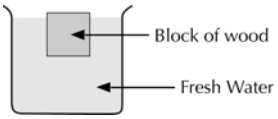
Hypothesis 2: When the mass of an object is less than its volume, the object will float in water.

Hypothesis 3: When an object's mass and volume are the same, the object will subsurface float in water.

**5b. Do the data in the table provide enough evidence to support all three hypotheses? Why or why not?**

***[INTERPRETATION-RULE, V]***

Use the following information to answer Questions 6a and 6b.




The picture shows a block of wood floating in fresh water. This same block will be placed in salt water.

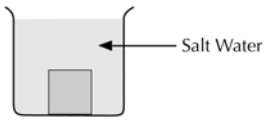
**6a. What do you know about sinking and floating that will help you decide what will happen when the block is placed in salt water?**

***[APPLICATION-RULE, CS/P]***

6b. Using your answer to 6a, which picture shows what will happen when this block is placed in salt water?

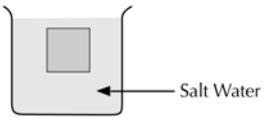
***[APPLICATION-CLAIM, V]***

A.



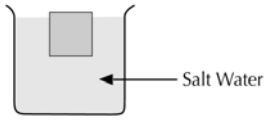

Sink

B.



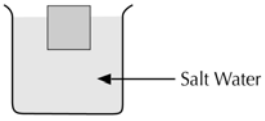

Subsurface Float

C.




Float

D.

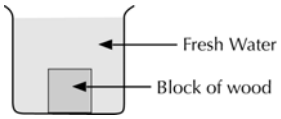



Float Higher

E.


Impossible to tell

Use the following information to answer Questions 6c and 6d.




The picture shows a block of wood sinking in fresh water. This same block will be placed in salt water.

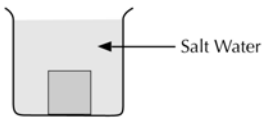
**6c. What do you know about sinking and floating that will help you decide what will happen when the block is placed in salt water?**

***[APPLICATION-RULE, CS/P]***

**6d. Using your answer to 6c, which picture shows what will happen when this block is placed in salt water?**

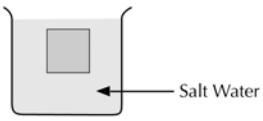
***[APPLICATION-CLAIM, V]***

A.



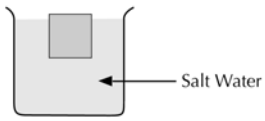

Sink

B.



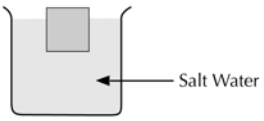

Subsurface Float

C.




Float

D.



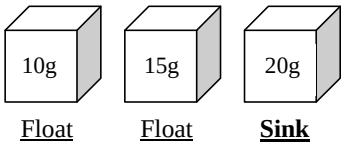

Float Higher

E.


Impossible to tell

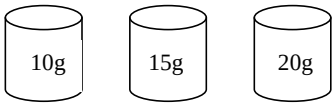
Use the following information to answer Questions 7a and 7b.

Jay placed the following objects in water and listed the results below each item.




Float Float **Sink**

All three cubes are of equal volume.




Float **Sink** **Sink**

All three cylinders are of equal volume.

Jay believes that the volume of the cubes is the same as the volume of the cylinders.

**7a. What do you know about sinking and floating that will help you decide whether Jay is correct?**

***[APPLICATION-RULE, 7a AND 7b SCORED TOGETHER, CS/P/V]***

**7b. Based on your answer to 7a, do you think Jay is correct? Why or why not?**
***[APPLICATION-CLAIM, 7a AND 7b SCORED TOGETHER, CS/P/V]***