A REPORT ON SCIENTIFIC REASONING BY UNDERGRADUATE STUDENTS AT THE UNIVERSITY OF VIRGINIA

Submitted by

The Undergraduate Scientific Reasoning Competency Assessment Committee

Committee Members:

Tony Baglioni       McIntire School of Commerce
Bobby Beamer        School of Continuing and Professional Studies
Jeanne Erickson    School of Nursing
Victor Luftig       Department of English, College of Arts and Sciences
Kirk Martini       School of Architecture
Aaron Mills        Department of Environmental Sciences, College of Arts and Sciences
Michael Palmer     Department of Chemistry, College of Arts and Sciences
Karen Schmidt      Department of Psychology, College of Arts and Sciences

July 2014

Coordinated by the Office of Institutional Assessment and Studies

Report written by Lois Myers and Sarah Schultz Robinson, IAS Staff
# Table of Contents

**EXECUTIVE SUMMARY** .............................................................................................................................. 1  
  Findings.................................................................................................................................................... 1  
  Recommendations............................................................................................................................... 2  
**SCIENTIFIC REASONING ASSESSMENT PLANNING** ........................................................................................ 3  
  Committee Charge.............................................................................................................................. 3  
  Definition of Scientific Reasoning..................................................................................................... 4  
  Student Learning Outcomes.............................................................................................................. 4  
  Standards/Level of Performance Expected .......................................................................................... 5  
  Background: Review of Scientific Reasoning Assessment Instruments....................................... 5  
**THE THREE ASSESSMENTS** .......................................................................................................................... 6  
  Planned Analyses.................................................................................................................................. 6  
**LONGITUDINAL ASSESSMENT OF LEARNING: ORIGINAL TEST** ................................................................. 7  
  Objectives............................................................................................................................................... 7  
  Methods................................................................................................................................................... 7  
  Results....................................................................................................................................................... 8  
  Shortening the Instrument and Modifying the Scoring Rubrics......................................................... 9  
**VALUE-ADDED AND COMPETENCY ASSESSMENTS: REVISED TEST** ....................................................... 10  
  Objectives............................................................................................................................................. 10  
  Methods................................................................................................................................................. 10  
  Results-Value-Added Assessment ........................................................................................................ 11  
  Results-Competency Assessment........................................................................................................ 14  
  Results-Test Validity............................................................................................................................. 16  
  Performance by Sub-Group .................................................................................................................. 16  
**FINDINGS** ............................................................................................................................................. 19  
**COMMITTEE RECOMMENDATIONS** ............................................................................................................ 20  
**APPENDICES** ......................................................................................................................................... 21  
**APPENDIX 1: ASSESSMENT COMMITTEE CHARGE AND MEMBERSHIP** .................................................. 21  
**APPENDIX 2: 2008-2009 REVIEW OF TESTS OF SCIENTIFIC REASONING** ........................................... 22  
**APPENDIX 3: THE 2010 SCIENTIFIC REASONING TEST AND ASSOCIATED RUBRICS** ....................... 24
EXECUTIVE SUMMARY

In 2012, the Office of Institutional Assessment and Studies initiated planning for a third assessment of undergraduate scientific reasoning competence, the previous assessments having occurred in 2004-2005 and 2009-2010. The Undergraduate Scientific Reasoning Competency Assessment Committee convened for this task built on the work of its predecessor, the 2009-2010 committee. That committee had created a definition of scientific reasoning, proposed learning outcomes and standards, designed a test to address the six learning outcomes, and administered it to first- and fourth-year students in 2010.

The 2009-2010 committee concluded that the 90-minute test provided a valid measure of students’ scientific reasoning competence and that students appeared to demonstrate competency in line with expectations. They subsequently recommended that the test be administered again in three years as a longitudinal assessment and that consideration be given to shortening the test so that it could be completed within 50 minutes.

The charge to the 2012-2013 Committee involved three simultaneous, inter-related objectives: improve the instrument, assess student competency, and assess student learning (value-added). To provide a comparative benchmark, especially to calibrate upper bound performance on the instrument by undergraduate students, the test was also administered to a sample of graduate students.

The assessments yielded evidence of competency and value-added in comparison to graduate students’ performance, especially clear for fourth-year students majoring in sciences. While on face value the test questions appeared to provide a valid measure of students’ scientific reasoning, the structure of the overall test may increase the likelihood of a Type II error (underestimating competence). Even graduate students, three-quarters of whom were from the sciences, achieved on average just over half of possible points.

Findings

1. Fourth-year students, on the whole, are capable of a respectable level of rigor in scientific reasoning.
2. Fourth-year students’ ability in scientific reasoning varies with field of study.
3. Evidence of value-added appears to be mixed, an observation that is likely a consequence of test limitations.
4. Graduate students’ test results serve well to calibrate the instrument and inform interpretation of undergraduate test results.
5. While individual test questions appear to provide a valid measure of students’ competence in scientific reasoning, the test overall may underestimate student competence.
6. The test illustrates the tension between the two aspects of assessment of student competency: breadth of the multiple outcomes assessed vs. depth of student knowledge and skill revealed for each outcome. A 50-minute test may not be able to serve both breadth and depth well.
Recommendations

1. Consideration should be given to the benefits and costs of revising and/or shortening the test even more, the purpose being to provide an opportunity for greater range and depth in student responses.

2. Analysis of students’ responses on each individual question should be undertaken to clarify which questions contribute most substantially to a robust understanding of students’ capabilities and which questions, if any, appear to provide redundant information.

3. Consistent with test improvements, the standards and scoring rubrics should be reviewed for validity.

4. Longitudinal studies should be designed and executed to assess progress from first to fourth year. The test or selected questions should be administered again in spring 2015-2016 to fourth-year students who took the test as first-years in 2012-2013.

5. In the next assessment of scientific reasoning, the test should be administered to another, much larger sample of graduate students to serve as a comparison group. The sample should include a diversity of disciplines in the sciences and social sciences, focusing on students in their 2nd to 4th years of graduate school.
SCIENTIFIC REASONING ASSESSMENT PLANNING

Committee Charge

The State Council of Higher Education for Virginia (SCHEV) mandates that all two- and four-year colleges and universities assess undergraduate competence in six core areas—written communication, quantitative reasoning, scientific reasoning, oral communication, critical thinking, and information literacy and technology (or an emerging area of interest can be substituted; UVa substituted undergraduate research).

In 2012-2013, the University focused on undergraduate competence in scientific reasoning, following up on the extensive groundwork laid by the assessment of scientific reasoning begun in 2008 and concluded in 2010. As the basis for that assessment, the Undergraduate Scientific Reasoning Competency Assessment Committee had defined scientific reasoning, posed learning outcomes, set expectations for student performance, critically reviewed available instruments, and then created and administered a new instrument to first- and fourth-year students. That assessment had yielded a set of findings and several recommendations, including three specific recommendations: 1) conduct a longitudinal assessment by re-testing in 2012 the students who had been tested as first-years in 2009, 2) shorten the test so that it could be administered in one hour or less, and 3) increase the sample size to provide for more in-depth analysis.

In 2012, a multi-disciplinary faculty committee, coordinated by staff of the Office of Institutional Assessment and Studies (IAS), initiated planning for the assessment. Committee membership represented a range of disciplines (Appendix 1). Four of the members had also served on the 2009-2010 committee.

The committee was asked:

- To review the 2009-2010 assessment of scientific reasoning
- To implement the report recommendations
- To conduct three assessments of scientific reasoning, including a longitudinal assessment, a cross-sectional assessment, and a competency assessment
- To interpret results, draw conclusions, and propose recommendations.

In reality, the charge entailed simultaneous assessment of:

1. The instrument and its associated rubrics and scoring—that is, how to fulfill the recommendation that the 90-minute test be shortened? On what basis should questions be eliminated? An evaluation of the instrument, question-by-question, was essential to address the recommendation to shorten the test.

2. Students’ competency in scientific reasoning
a. While several instruments are available to assess basic (e.g., high school or first-year college) students’ ability, a wide search found no established tests aimed at higher level scientific reasoning that offered evidence or results for comparison.

b. Assuming use of a shortened instrument, what standard could be applied to assess undergraduate student performance? Testing of graduate students in the sciences would provide results for comparison, in essence a standard that undergraduates would not likely exceed.

3. Student change in competency (value-added) Both longitudinal and cross-sectional designs were applied to assess differences in scientific reasoning skills between first-year and fourth-year students.

This three-part charge determined the following schedule:

- **Fall 2012:** Longitudinal value-added assessment (original instrument)
- **Winter 2012-13:** Evaluation of original instrument in light of longitudinal assessment results
  - Revision/shortening of the instrument; revision of rubrics
- **Spring 2013:** Administration of the revised instrument to assess learning and competency (value added, cross-sectional designs), including administration to graduate students

This report describes the results for each of the three related components of the overall assessment. Results of the assessment will be used to satisfy SCHEV assessment requirements and to provide information that would be useful for faculty, deans and the Provost as they make curricular and resource decisions.

**Definition of Scientific Reasoning**

Scientific reasoning is a mode of thought that:

- draws on systematic observation and description of phenomena;
- employs established facts, theories, and methods to analyze such phenomena;
- draws inferences and frames hypotheses consistent with that body of public knowledge and understanding;
- subjects explanations to empirical tests, including scrutiny of their declared and latent assumptions; and
- allows the possibility of changes in explanations as new evidence emerges.

**Student Learning Outcomes**

UVA expects all undergraduates to be able to employ scientific reasoning for their own purposes but especially for the purpose of evaluating the quality of scientific information, argument, and conclusions. Graduating fourth-year students at the University of Virginia are expected:

1) To understand that, while scientific statements are in principle tentative, criteria exist by which they can be judged, including consistency with the body of scientific theory, method, and knowledge;

2) To display a grasp of experimental and non-experimental research design, including the notion of control, the idea of statistical significance, and the difference between causation and association;
3) To interpret quantitative data presented in graphical form;
4) To acknowledge the possibility of alternative accounts of events and properties and judge their relative plausibility by standard criteria;
5) To identify sources of error in scientific investigation, including errors of measurement and ambiguity of judgment;
6) To detect unsound logic in scientific arguments.

**Standards/Level of Performance Expected**

- 25% of students highly competent (score 71-100)
- 75% competent or above (score 56-70)
- 90% minimally competent or above (score 46-55)
- 10% not competent (score 45 or lower)

**Background: Review of Scientific Reasoning Assessment Instruments**

After defining scientific reasoning and proposing associated learning outcomes, the committee then evaluated available instruments in light of the definition and learning outcomes (see Appendix 2 for review of instruments as described in 2010 committee report).

In response to the dearth of appropriate instruments, the 2009-2010 committee created a new instrument that would address the set of learning outcomes. The test employed excerpts from scientific publications as the basis for three types of test questions: multiple choice, short essay, and experimental design essay. After pilot-testing the instrument, IAS administered the test to a sample of first- and fourth-year students, and committee members scored the essay answers according to rubrics. The committee concluded that the test appeared to provide a valid measure of students’ competence in scientific reasoning, that the test was more thorough and challenging than the CAAP test in the 2004-2005 assessment, but that the instrument was too long (90 minutes) and needed trimming.

Seeking to discover if an adequate test of scientific reasoning had emerged in the intervening years, the 2012-2013 committee considered results from a query posed to AAU institutions regarding use of tests of scientific reasoning. Most commonly mentioned—James Madison University’s *Scientific Reasoning Test* and the *Lawson Classroom Test of Scientific Reasoning*—were described as elementary, perhaps useful for first-year students, but not adequate to assess reasoning skills of fourth-year students. In fact, Princeton University, citing a similar unsuccessful quest for a scientific reasoning test that addressed higher order thinking, requested the opportunity to see the instrument designed at U.Va. The other instruments that were suggested either were too new/unknown or were focused too narrowly. Without finding a suitable alternative, the committee proceeded to conduct the assessment with the UVa instrument.

---

1. CLASS—Colorado Learning Attitudes about Science Survey—student attitudes about learning physics
   TOSLS—Test of Scientific Literacy Skills—multiple choice, used in general education courses (non-bio-majors)
   EDAT—Experimental Design Ability Test—focused on experimental design
   CLA+—Collegiate Learning Assessment Test—90 minute test, formerly the CLA, added questions re: quantitative and scientific reasoning (one sub-score). The CLA had not been administered at UVa as it was found to have a “ceiling effect” at comparable institutions. No information is available regarding a ceiling effect for the CLA+.
THE THREE ASSESSMENTS

Planned Analyses
The overall assessment in 2012-2013 entailed three analyses that built on the results of the 2009-2010 assessment. Figure 1 portrays the analyses of value-added and competency conducted for both the 2009-2010 and the 2012-2013 assessments.

#1. **Longitudinal (Value-added):** Using the original test and a longitudinal design, comparison of fourth-year students’ performance to their performance three years earlier as first-year students. This assessment, which was requested by the 2009-2010 committee, would inform decisions about test revision.

#2. **Cross-sectional (Value-Added):** Using the revised test and a cross-sectional design, comparison of first-year and fourth-year students’ results to assess value-added.

#3. **Competency:** Using the revised test and comparison to graduate student performance, assessment of fourth-year students’ competency.

---

**Figure 1: Scientific Reasoning Assessment Comparisons: 2009-2010 and 2012-2013**

<table>
<thead>
<tr>
<th></th>
<th>2009-2010 Assessments</th>
<th>2012-2013 Assessments</th>
<th>Grad students</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st years</td>
<td>4th years</td>
<td>1st years</td>
</tr>
<tr>
<td>Longitudinal (Value-added)</td>
<td>Original 90 min test</td>
<td>#1</td>
<td>Original 90 min test</td>
</tr>
<tr>
<td>Cross-sectional (Value-added)</td>
<td>Original 90 min test</td>
<td>Original 90 min test</td>
<td>Revised 50 min test</td>
</tr>
<tr>
<td>Competency</td>
<td>Original 90 minute test; Compare to standard</td>
<td>Revised 50 min test #3</td>
<td>Revised 50 min test</td>
</tr>
</tbody>
</table>
LONGITUDINAL ASSESSMENT OF LEARNING: ORIGINAL TEST

Objectives

- Assess change in performance within a sample of fourth-year students, comparing their performance as fourth-years with their performance as first-year students
- Assess value of each question in distinguishing first-year and fourth-year performance

Methods

Test

The same 90-minute test that was administered in 2009-2010 was used in the Fall 2012 assessment, and the same rubrics were applied to score student responses (see Appendix 3 for test and rubrics).

Sampling

Sixty-six fourth-year students who had taken the test as first-years were invited by email to participate in an assessment and were offered gift cards as compensation. While informed that they were specifically invited because they had taken a test as first-years, they were not informed of the topic of the assessment. In total, 35 fourth-year students (21 women and 14 men) volunteered to participate in the assessment. One student was from Architecture, 24 from the College of Arts and Sciences (8 Humanities/Fine Arts, 3 Social Sciences, 13 Sciences), three from the McIntire School of Commerce, six from the School of Engineering and Applied Science, and one from the School of Nursing. Compared to the student population and to the sample in the 2009-2010 assessment, the sample was “science-heavy.”

Administration and Scoring

Students who consented to take the test were assured of confidentiality. Participants were given a $50 gift certificate to Amazon.com and snacks at the test sessions, which were administered by IAS staff in October-November 2012. After the sessions, IAS staff recorded scores for the multiple choice questions and convened the scoring sessions for the short essay and experimental design answers.

In scoring sessions, the team of scorers applied the rubrics to each of the short answer and experimental design questions after norming for each question. They assigned scores of Not Competent=1, Marginally Competent=2, Competent=3, and Highly Competent=4 to each component of the rubric (corresponding to a learning outcome) for each question. Each answer received two readings. IAS staff conducted data entry and analysis of the rubric results.

About 90 percent of independent ratings agreed within one point. Scorers’ ratings matched exactly 44 percent of the time, forty-five percent of ratings differed by one point, and 11 percent differed by more than one point. These results fall within the range of acceptable reliability.
Results

Student Performance

Students’ performance on the test as fourth-years was compared to their prior performance on the test as first-years. The difference between the mean total scores for first-years and fourth-years was not statistically significantly different, although it approached significance ($p=.08$). Fourth-years earned on average 66.9 percent of possible points, compared to 62.9 percent of possible points as first-years (Table 1). On the whole, however, results based on total scores were inconclusive, in part due to the small sample.

Table 1: Longitudinal Assessment- Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Years</td>
<td>35</td>
<td>20.63</td>
<td>46.22</td>
<td>36.61</td>
<td>6.22</td>
</tr>
<tr>
<td>4th Years</td>
<td>35</td>
<td>30.36</td>
<td>49.86</td>
<td>39.45</td>
<td>4.67</td>
</tr>
</tbody>
</table>

In general, fourth-year students’ scores were significantly higher than their first-year scores on short-answer questions ($p<.05$) but not on multiple choice questions or on experimental design questions (Figure 2). Comparison of the mean scores by question indicated that three of the five short answer questions showed statistical differences between first- and fourth-years ($p < .05$ and $p < .09$), and for only one of eleven multiple choice question was there a statistical difference for fourth-year performance over first-year, and neither experimental design question showed any statistical differences between first- and fourth-years.

Figure 2: Mean Total Scores by Type of Question: Scores in First-Year vs. Fourth-Year
Shortening the Instrument and Modifying the Scoring Rubrics

The task of shortening the test relied on evaluating the degree to which scores on each question discriminated levels of competence (i.e., high performers from low performers):

**Multiple choice questions:** On the whole (with the exception of one question—Q1), performance on the multiple choice questions did not differ between the students’ first and fourth years, suggesting that multiple choice questions did not contribute to the discrimination of competence and thus were likely candidates for deletion from the test (Appendix 4: Table 1 and Figure 1).

**Short-answer questions:** Scores on three of the short answer questions (Q4, Q11, and Q15) showed the greatest ability to discriminate between students’ performance in the first year and their performance in their fourth year (Appendix 4: Table 1 and Figure 3).

**Experimental design questions:** Although differences between students’ performance in their first year and their fourth year were not statistically significant for either of the two experimental design questions, the committee was loath to eliminate design questions from the test. Students’ ability to design an experiment was seen as a best indicator of higher level scientific reasoning even though students may not have learned about experimental design during their undergraduate tenure (Appendix 4: Table 1 and Figure 2).

Based on these test results and on the scorers’ experience in grading the answers, the committee recommended eight changes to shorten and improve the instrument:

- Eliminate all multiple choice questions
- Keep two design questions, but replace one experimental design question with an observational design question
- Eliminate and replace selected short-answer questions
- Create uniform space for students to write short answers
- Edit advice to students regarding how much time they should spend writing each answer
- Forego one learning outcome (*Detects unsound logic in scientific arguments*) as it would no longer be addressed by the remaining questions
- Administer the test through online survey software in group testing settings so that students type their answers, thereby alleviating the challenge of reading students’ handwritten answers
- Revise scoring rubrics so that scorers do not assess each student’s short answer according to multiple learning outcomes. It was felt that the short answers were too limited in length and depth to “carry the weight” of multiple separate outcomes assessments.

IAS revised the test and rubrics in accordance with the recommendations and in preparation for the subsequent assessments. While the original 90-minute test contained 18 questions for scoring (11 multiple-choice, 5 short-answer, and 2 experimental design essay questions), the revised (estimated 50-minute test) included eight questions—six short-answer and two design questions (See Appendix 5 for revised instrument and rubrics.). Evaluators mapped each of the remaining eight questions to each learning
outcome (Appendix 6). One outcome—detect unsound logic in scientific arguments—was no longer represented by a test question and so was dropped.

VALUE-ADDED AND COMPETENCY ASSESSMENTS: REVISED TEST

Objectives
• Assess value-added by comparing first-year and fourth-year student performance (cross-sectional design) for a sample of students
• Assess fourth-year students’ competency in scientific reasoning by comparing their performance with that of graduate students, which will serve as the standard for highest possible performance
• Analyze differences in performance by discipline
• Analyze differences in performance by test question and question type

Methods

Test
The revised, 8-question test was administered in spring 2013, and the revised rubrics were applied to score student responses (see Appendix 5 for revised instrument).

Sampling
In spring 2013, a stratified random sample of 1,411 first-year and 1,522 fourth-year students was invited by email by the Vice Provost for Academic Affairs to participate in the assessment. The invitation did not specify the topic of the assessment. The sample of fourth-years was stratified by gender, school, and discipline for those in the College of Arts and Sciences. Seventeen percent (504 students) of those invited responded to the request. Of these, 410 were able to be scheduled to attend a testing session. Of the 410 students scheduled, 340 attended and took the assessment test (17% no-show rate).

In addition, committee members in the sciences and in the School of Architecture invited graduate students to take the test. In all, 27 graduate students submitted completed tests. Of the 27, seven were first-years, seven second-years, four third-years, and nine fourth-to-sixth years. Fifty-two percent were from Psychology, 26 percent from Environmental Sciences, and 22 percent from Architecture. In total, 367 students completed the test: 170 first-years, 170 fourth-years, and 27 graduate students.

The sample of first- and fourth-years also was compared to the population to evaluate differences in discipline, gender, or measures of academic potential and success: current GPA, and SAT Verbal, Math, and Writing scores (Appendix 7). Prior to analysis of the results, the undergraduate samples were corrected for overrepresentation by some disciplines by randomly selecting students’ tests to be deleted. Female students were over-represented in the sample, but this overrepresentation was not considered to be of serious concern as gender differences were not revealed in performance on any test question. Also, among fourth-years, the sample showed statistically significantly higher GPA and SAT scores, although effect sizes were small (Appendix 7).
Final samples for analysis included 109 first-year students and 156 fourth-year students.

**Test Administration**

Students who took the test were assured of confidentiality. All students who took the test, including those who were not able to finish it, received a $20 gift certificate to Amazon.com in thanks for their participation.

About 40 percent of undergraduate students did not answer all eight questions. Eighty-seven percent made it through Question #7 but only 64 percent answered Question #8. Among graduate students taking the test, 11 percent did not answer all eight questions, with seven percent not answering Question #8. Some students skipped questions, probably with the intention of returning later, but ran out of time. In cases where a student did not complete the last or second-to-last question, a score was not included in the analysis for that question and the overall scoring was adjusted for the exclusion.

**Scoring Reliability**

Fifteen faculty and graduate students scored the scientific reasoning tests over a four-day workshop in May 2013. Raters were unaware of which tests had been completed by first-years, fourth-years, or graduate students. Papers were scored according to rubrics, and norming sessions were conducted prior to scoring. Each short answer question was scored on a scale of one to six, and the experimental design questions were scored on a scale from one to twelve (See Appendix 5 for test and rubrics).

Scoring reliability is considered acceptable if no more than ten percent of ratings by evaluators differ by more than one point where the scoring range is one-to-four. In the assessment of short answer questions (scoring range 1 to 6), 14 percent differed by more than one point on a six-point scale (11 percent by two points and 3 percent by more than two points). Most ratings were either equal or differed by just one point (Appendix 8). The scoring range for the experimental design questions was twice that of the short-answer questions (1 to 12). Applying the same standard but simply doubling it results in acceptable reliability if not more than ten percent of ratings differ by more than two points. The scoring of the experimental design questions met this standard. Ten percent of ratings differed by more than two points, 27% were equal, and 63% differed by one or two points (Appendix 8).

**Results-Value-Added Assessment**

In this cross-sectional design, the sample of fourth-year students’ mean scores was compared to mean scores from the sample of first-year students. Overall, performance on the scientific reasoning test did not differ significantly between first- and fourth-year students for any questions (Table 2).
Table 2: Short Answer Questions-Descriptive Statistics

<table>
<thead>
<tr>
<th>Question</th>
<th>First-Year (n=109)</th>
<th>Fourth-Year (n=156)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Q1- Criteria to judge scientific statements</td>
<td>2.13</td>
<td>1.23</td>
</tr>
<tr>
<td>Q2- Criteria to judge scientific statements</td>
<td>3.25</td>
<td>1.13</td>
</tr>
<tr>
<td>Q3- Experimental Design</td>
<td>4.66</td>
<td>1.86</td>
</tr>
<tr>
<td>Q4- Alternative Accounts</td>
<td>2.56</td>
<td>1.12</td>
</tr>
<tr>
<td>Q5- Sources of Error</td>
<td>2.47</td>
<td>1.41</td>
</tr>
<tr>
<td>Q6- Graph</td>
<td>3.12</td>
<td>1.53</td>
</tr>
<tr>
<td>Q7- Alternative Accounts</td>
<td>2.71</td>
<td>1.07</td>
</tr>
<tr>
<td>Q8- Experimental Design</td>
<td>4.18</td>
<td>1.69</td>
</tr>
</tbody>
</table>

Fourth-year science majors, however, clearly outperformed first-year students on two questions, approaching significance on two additional questions. Fourth-year science majors were more likely to score higher on Q5 (Sources of Error) (p<.05) and on Q3, the first Experimental Design question (p<.01) (Figures 3 and 4). Also suggestive of differences were results for two more questions (Q6-Graphs and Q7-Alternative Accounts). Except for the means for Questions 2 and 4, which were almost identical, fourth-year science majors tended to score higher than first-year students.
Figure 3: Mean Scores for Fourth-Year Science Majors vs. First-Year Students on Short Answer Questions

Figure 4: Mean Scores for Fourth-Year Science Majors vs. First-Year Students on Experimental Design Questions
Overall, these results suggest that among science majors there is evidence of improved performance over time. Among the general population of fourth-year students, however, the evidence of increased ability to reason scientifically, especially with regard to experimental design, is absent.

**Results-Competency Assessment**

On average, fourth-year students earned only about 45 percent of total possible points for questions answered. Taken at face value, these results might suggest that as a group fourth-years did not perform well on the test. The performance of the 27 graduate students, however, provides a useful comparison: graduate students, three-quarters of whom are in the sciences, achieved only 54 percent of possible points. For a demanding test, a score of 54 percent may well be a reasonable expectation.

We readily acknowledge that the sample of graduate students is too small to sustain a firm conclusion. If we make the assumption, however, that the sample of graduate students does represent the population of “ideal responders,” it would then follow that on average fourth-year students demonstrate competency in scientific reasoning (Figure 5). Indeed, fourth-year students’ mean scores matched graduate students’ for two questions (Q2 and Q7), were within 90 percent of graduate students’ for two more questions (Q1 and Q6), and within 75 percent for three more questions (Q4, Q5, and Q8). The two questions for which there was the greatest disparity between fourth-years and graduate students’ performance—Q3 and Q8—were both experimental design questions that required higher level, creative thinking.

The original standards for performance had proposed that highly competent students would earn scores of 71 percent or more and competent students 56 percent or more. Applying this standard to these results, we could surmise that in comparison to the standard set by the sample of graduate students, fourth-year students exhibited competence in scientific reasoning.

**Figure 5: Percentage of Points Earned by Question with Maximum=Graduate Student Performance.**

Examining the results by question, graduate students’ mean scores were significantly higher than for first- and fourth-year students for five questions: three short answer questions and both experimental design questions (Figures 6 and 7). Among the short answer questions, graduate students scored significantly higher than first- or fourth-year students on Question 4 (Alternative Accounts), Question 5 (Sources of
Error), and Question 6 (Graphs). For three questions (Q1, Q2, Q7), differences in performance did not rise to the level of statistical significance. The frequency distributions for first-year, fourth-year and graduate student performance for each question can be found in Appendix 9.

**Figure 6: Comparison of First-Year, Fourth-Year, and Graduate Student Performance: Short Answer Questions**

![Figure 6: Comparison of First-Year, Fourth-Year, and Graduate Student Performance: Short Answer Questions](image)

**Figure 7: Comparison of First-year, Fourth-year, and Graduate Student Performance: Experimental Design Questions**

![Figure 7: Comparison of First-year, Fourth-year, and Graduate Student Performance: Experimental Design Questions](image)
Results-Test Validity

If we apply the original competency standards to the 27 graduate students who completed the test, the results would suggest that either the graduate students (nearly all of whom are in science disciplines) do not meet expectations for competency or test validity is an issue (Table 3).

Table 3: Graduate Student Performance compared to Standards for Undergraduate Performance

<table>
<thead>
<tr>
<th>Percentage of Total Points Earned</th>
<th>Graduate Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly competent</td>
<td>At least 71%</td>
</tr>
<tr>
<td>Competent or better</td>
<td>At least 56%</td>
</tr>
<tr>
<td>Minimally competent or better</td>
<td>At least 46%</td>
</tr>
<tr>
<td>Not competent</td>
<td>Less than 46%</td>
</tr>
<tr>
<td>Number</td>
<td>1</td>
</tr>
<tr>
<td>Percent</td>
<td>4%</td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td></td>
</tr>
</tbody>
</table>

Based on these results and the assumption that the 27 graduate students do indeed represent the population, it would be reasonable to posit that, even with the changes made to shorten the test:

- The test topics and questions are very difficult.
- Too little time is allotted for students of any level to demonstrate competence.
- Scoring standards are too stringent.

Any, all, or combinations of the above may be a factor in the results that emerged.

A comparison of three questions for which the text remained the same between the original and revised test suggests that the change in time allotted, and possibly the change in scoring, may have dampened the 2013 test’s ability to detect levels of competence (Appendix 10).

Performance by Sub-Group

Fourth-year students’ performance on each question varied according to school and discipline (Figures 8 and 9). Science, Social Science, and Commerce majors significantly outperformed Humanities/Fine Arts and Engineering majors on the short answer question on Sources of Error. Science majors performed the best on the first Experimental Design question (Q3), scoring significantly higher than Humanities/Fine Arts majors (p<.05).
Within the College, the differences in performance were as expected: for six of the eight questions, social science and science majors scored similarly (Figures 10 and 11). Science and Social Science majors significantly outperformed Humanities/Fine Arts majors on Q5 (Sources of Error), and Science majors significantly outperformed Humanities/Fine Arts majors on Q3 (Experimental Design).
Comparisons among the subgroups provide additional evidence that science majors are more likely to exhibit equal or greater levels of scientific reasoning than their colleagues in other disciplines. Their predominance is not universal or compelling, however. Students majoring in the social sciences and in Commerce especially also show considerable skill across multiple test questions, while Engineering students did not.
FINDINGS

1. **Fourth-year students, on the whole, appear to be capable of a respectable level of rigor in scientific reasoning.** Fourth-year UVa students appear to be competent in scientific reasoning as judged by comparison with graduate student performance. Students appear to be able to draw sound inferences from graphic presentations of quantitative information, recognize sources of error, and know that plausible alternative scientific hypotheses need systematic tests. UVa fourth-year students appear, by and large, to be well versed in experimental design and able to apply design principles to open-ended problems.

2. **Fourth-year students’ ability in scientific reasoning varies with field of study.** There is a significant difference in performance between fourth-year students in the College of Arts and Sciences who have studied science and mathematics and those who have studied humanities and fine arts, with the science majors outperforming others. Students majoring in the social sciences and in Commerce also demonstrate considerable skills in scientific reasoning.

3. **Evidence of value-added is not strong, except with regard to natural science/math majors in the College of Arts and Sciences.** Science/math fourth-years outperformed first-years on multiple questions on the scientific reasoning assessment, although fourth-years overall did not outperform first-years.

4. **Graduate students’ test results appear to serve well to calibrate the instrument and inform interpretation of the undergraduate test results.** The sample of graduate students was too small, however, to sustain rigorous analysis or to support firm conclusions.

5. **While the test questions, which address aspects of scientific reasoning presented through articles published in scientific publications, provide evidence of face validity of the test, the overall test appears to limit the demonstration of competency among both undergraduate and graduate students.** The test appears to provide too little time for students to be able to both complete the test and demonstrate their competence. Students given more time score higher. Additional changes are needed to allow for the full range of competency to be demonstrated.

6. **The test illustrates the tension between the two aspects of assessment of student competency: breadth (multiple learning outcomes that represent the important aspects of scientific reasoning) and depth (range of student skills and abilities from first-year students through fourth-year non-science and science majors).** A 50-minute test may not be able to serve both aspects well.
COMMITTEE RECOMMENDATIONS

1. **Consideration should be given to the benefits and costs of revising and/or shortening the test even more, the purpose being to provide an opportunity for greater range and depth in student responses.** In its present form, the test may reward quick thinking and penalize students for whom the concepts do not come immediately to mind or who need time to deliberate, whether or not they major in the sciences. In contemplating a shorter test, however, it would be vitally important to maintain the strong features of the test and to ensure that the information obtained would be as reliable and useful in spite of the reduction in data obtained. Both committees valued the test for three features: it does not require prior knowledge of facts in any specific field of science; it calls for written responses beyond simple choice among alternatives; and it gives an opportunity for creative responses. Test revisions should be designed to support these three features.

2. **Analysis of students’ responses on each individual question (Appendix 9) should be undertaken to clarify which questions contribute most substantially to a robust understanding of students’ capabilities and which questions, if any, appear to provide redundant information.** Specific recommendations:
   a. Delete questions 2 and 7
   b. Option to consider: split test so that students randomly receive a sample of questions, not all questions, e.g., half of sample takes one half of the test, etc..
   c. Include a multiple choice question as the first question—to help orient students to the task quickly
   d. Add a question that probes whether students perceive that they had sufficient time to complete the test adequately.

3. **Consistent with test improvements, the standards and scoring rubrics should be reviewed for validity.** Given the amount of time allotted per question, are the scoring standards appropriate to the task?

4. **Longitudinal studies should be designed and executed to assess progress from first to fourth year.** The test or selected questions should be administered again in spring 2015-2016 to fourth-year students who took the test as first-years in 2012-2013.

5. **In the next assessment of scientific reasoning, the test should be administered to another, much larger sample of graduate students to serve as a comparison group.** The sample should include a diversity of disciplines in the sciences and social sciences, focusing on students in their second to fourth years of graduate school.
Appendices

Appendix 1: Assessment Committee Charge and Membership

**Background:** The State Council of Higher Education for Virginia mandates that all of its two and four year colleges and universities assess undergraduate competence in six core areas—writing, quantitative reasoning, scientific reasoning, oral communication, critical thinking, and information literacy and technology (or an emerging area of interest can be substituted; UVa substituted undergraduate research). In the past, SCHEV has required the assessment of five percent of graduating fourth-years. However, the University has always exceeded this requirement in order to analyze results by school—where curricular decisions are made. While the ostensible purpose of the core competency assessments is for students and parents to be able to compare institutions based on student achievement, the actual genesis of the effort lay in the state legislature’s response to anecdotal reports that graduating students were not competent in basic areas. Given the mandate, the University has designed assessments that support its own purposes and also meet SCHEV requirements. Useful information has come out of the assessments to assist faculty, deans and the Provost in improving student learning in these core areas. To learn more about core competency assessment, go to: [http://avillage.web.virginia.edu/iaas/assess/data/competency.shtm](http://avillage.web.virginia.edu/iaas/assess/data/competency.shtm)

**Scientific Reasoning Assessment Committee:** All committee work will be facilitated by the Office of Institutional Assessment and Studies. Logistics, data collection and analysis, and distribution and replication of materials will be managed by IAS. As the modus operandi, faculty do the cognitive, evaluation, and analysis work, and IAS staff do the facilitative, logistical, support and coordination work.

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baglioni, Tony</td>
<td>McIntire School of Commerce</td>
</tr>
<tr>
<td>Beamer, Bobby</td>
<td>School of Continuing and Professional Development</td>
</tr>
<tr>
<td>Erickson, Jeanne</td>
<td>School of Nursing</td>
</tr>
<tr>
<td>Luftig, Victor</td>
<td>English department</td>
</tr>
<tr>
<td>Martini, Kirk</td>
<td>School of Architecture</td>
</tr>
<tr>
<td>Mills, Aaron</td>
<td>Environmental Sciences department</td>
</tr>
<tr>
<td>Myers, Lois</td>
<td>Institutional Assessment and Studies</td>
</tr>
<tr>
<td>Palmer, Michael</td>
<td>Chemistry department and TRC</td>
</tr>
<tr>
<td>Schmidt, Karen</td>
<td>Psychology department</td>
</tr>
<tr>
<td>Schultz Robinson, Sarah</td>
<td>Institutional Assessment and Studies</td>
</tr>
</tbody>
</table>

Collegiate Assessment of Academic Proficiency

The group reviewed the previous exercise in assessment of UVa students’ capability in scientific reasoning, which consisted of administration and (limited) interpretation of an external, nationally-normed standardized test. This was the Collegiate Assessment of Academic Proficiency developed by the ACT (American College Testing) corporation. According to the description at the ACT web site, CAAP is designed to emphasize “scientific reasoning skills rather than recall of scientific content or a high level of skill in mathematics or reading” and is composed of forty-five multiple choice questions on content drawn from biology, chemistry, physics, and physical sciences. Questions deal with data representation, design and interpretation of experiments, and interpretation of conflicting hypotheses.

The University’s experience with the CAAP in the prior assessment of scientific reasoning produced some informative results. The test was administered to a five percent sample of fourth-year students. Science majors did very well on the test, scoring in the 94th percentile. The rest of the College of Arts and Sciences scored in the 82nd percentile. The School of Commerce scored in the 87th percentile. The Schools of Architecture and Education scored a bit lower, in the 76th and 71st percentiles, respectively. The limitations of this exercise are evident. It is plausible that science majors, who have acquired specific scientific information and who presumably have had practice with the forms of scientific reasoning, should do better than those for whom scientific thought is less central to their interests and preoccupations. The ordering of results for other groups is likewise unsurprising. But was there is a clear hint of ways to improve the power and subtlety of our students’ reasoning?

A copy of a previous CAAP test was made available to the committee chair (with confidentiality enforced). The care with which the test was constructed was evident. We praise the worthy intent and the careful validation of the multiple-choice questions. The design of the test comports in broad outline with our purposes. In the judgment of the reviewers, however, it appeared that the test was far from free from the need to recall specific technical information. Members suspected that multiple choice questions would necessarily suppress the kind of information that written answers could convey. Furthermore the brief time allowed for the test

---

2 http://www.act.org/caap/test_science.html
3 The test consists of eight passage sets, each of which contains scientific information and a set of multiple-choice test questions. A passage may conform to one of the three different formats listed below.

- **Data Representation.** This format presents students with graphic and tabular material similar to that found in science journals and texts. The items associated with this format measure skills such as graph reading, interpretation of scatterplots, and interpretation of information presented in tables, diagrams, and figures.

- **Research Summaries.** This format provides students with descriptions of one experiment or of several related experiments. The items focus on the design of experiments and the interpretation of experimental results. The stimulus and items are written expressly for the Science Test, and all relevant information is completely presented in the text of the stimulus or in the test questions.

- **Conflicting Viewpoints.** This format presents students with several hypotheses or views that are mutually inconsistent owing to differing premises, incomplete or disputed data, or differing interpretations of data. The stimuli may include illustrative charts, graphs, tables, diagrams, or figures. Items in this format measure students' skills in understanding, analyzing, and comparing alternative viewpoints or hypotheses.
(40 minutes for 45 questions) seemed to limit the depth of the response. Perhaps more significant, it seemed difficult to express how each question bore on the five criteria for ability in scientific reasoning that the committee had established.

**James Madison University’s Scientific Reasoning Test**

James Madison University’s ambitious assessment effort includes a carefully developed and validated scientific reasoning test. It is “recommended for use with general education program assessment… for undergraduate college students.” Items in this multiple choice, on-line test are “designed to be content-free” and to address learning objectives that include description of scientific methods of inquiry, design of meaningful experiments, and use of theories and models to explain natural phenomena. Considerable emphasis is given to the use of scientific information to address public issues. A test manual is available, as well as a short version of the test itself (10 questions of the 49 in the full instrument). The test’s reliability, expressed technically by Cronbach’s alpha, meets the research standard (values of α exceed 0.7). Judgment of validity is more qualitative; according to the manual, the test’s validity was strengthened by an explicit mapping from each item to one or more of the stated objectives.

The committee chose not to pursue use of the JMU test for several reasons. The multiple-choice format seemed incomplete; members thought they could see more deeply into students’ reasoning by reading written short answers and essays. Members had some concern that logistics of test administration would be complex. Cost was an issue. Finally, inspection of the brief sample of test items suggested that their emphasis on the social dimensions of science would lead the assessment away from the objectives established. Instead, the committee resolved to try its hand at constructing a test consistent with members' ideas of scientific reasoning and expectations for quality of responses.

**The Collegiate Learning Assessment**

Very recently the Collegiate Learning Assessment (CLA) has received considerable attention, as it is the foundation for a significant part of the research described in Arum and Roksa’s new book *Academically Adrift*. This important book has much to say on the impact of the college years on students’ ability to reason and to write. In brief, the CLA asks for analysis of complex situations and a written proposal for action. The value of such a test is clear; however, its relevance to assessment of scientific reasoning is indirect. The committee did not include the CLA among possible instruments for the scientific reasoning assessment.

---

Appendix 3:  The 2010 Scientific Reasoning Test and Associated Rubrics

UNIVERSITY OF VIRGINIA CORE COMPETENCY ASSESSMENT
Scientific Reasoning 2010

Instructions:

Please do not begin until the IAS staff person has given you instructions to do so. Once instructed to begin, you will have 90 minutes to complete the assessment.

This instrument was developed by University faculty and administrators, with input from students, and is being administered for the sole purpose of improving undergraduate education. Your score will never be a part of your academic record. It will be used to compute composite scores for the University, your school, and perhaps your major.

The goal of this assessment is to help the University better understand how well our students can reason using scientific principles and methods. We want to assess how well the University is teaching science reasoning to all undergraduates, not just science majors. Many aspects of scientific reasoning are covered herein, and some questions are more challenging than others. While it is important to answer as many questions correctly as you can, PLEASE DO NOT GUESS RANDOMLY. If you run out of time, or have no idea of an answer, just leave the question blank. Random guessing will detract from the usefulness of the data we receive. We want to know what you know and can do—whether that is a little or a lot.

The assessment contains twelve multiple-choice and five short answer questions, and asks you to design two experiments. Please attempt to answer all the short answer and experimental design questions (Questions 3, 4, 5, 7, 11, 15, 18). For these written components, please keep in mind that we are assessing scientific reasoning skills, not writing ability. While faculty evaluators must be able to understand what you have written, grammar, punctuation, style and syntax will NOT be assessed. You may use bullet points and abbreviations.

The assessment was designed to be taken in the order the questions appear. You may skip difficult questions though, and return to them as you wish.

Some of the passages have been edited for the purposes of designing this instrument. In cases where the passage has been edited, ellipses (...) or brackets [ ] denote removed or added text, respectively.

Thank you for your participation!

COPYRIGHTED MATERIALS INCLUDED
REPRINTED FOR RESEARCH PURPOSES ONLY
NO FURTHER DISTRIBUTION

---

9 Space supplied for answers by students has been removed from the following pages.
Institutional Assessment and Studies


**Passage 1**: The ability to think and plan is taken by many of us to be the hallmark of the human mind. Reason, which makes thinking possible, is often said to be uniquely human and thus sets us apart from the beasts. In the past two decades, however, this comfortable assumption of intellectual superiority has come under increasingly skeptical scrutiny. Most researchers now at least entertain the once heretical possibility that some animals can indeed think.

In an early, controversial study (1914), German psychologist Wolfgang Köhler was working at a primate research center on the Canary Islands. He presented novel problems to chimps he had caught in the wild. Often the pattern of solution suggested insight rather than trial and error. For instance, when Köhler first hung a bunch of bananas out of reach, the chimpanzee being observed made a few useless leaps, then went off to a corner and "sulked." But in time the chimp looked back at the bananas, then around the large outdoor enclosure at the various objects he had to play with, back to the bananas, back to one specific toy (a box), then ran directly to the box, dragged it under the fruit, climbed on top, leaped up and grabbed the prize.

**Question 1.** Köhler argued that his observations lent support to a hypothesis that some animals are able to plan actions in advance. What alternative explanation could reasonably account for Köhler's observations?

A. In the wild, Köhler's chimps use a wide variety of methods to gather food.
B. In the wild, Köhler's chimps climb trees to harvest bananas.
C. In the wild, Köhler's chimps move and stand on objects to reach food.
D. In the wild, Köhler's chimps jump from branch to branch to get to food.
E. All of the above.

**Question 2.** Which experimental scenario would strengthen Köhler's conjecture that chimps are able to conceive of a plan to get bananas?

A. A lab-raised chimp who has never seen the banana problem is placed in an enclosure. An assortment of objects familiar to the chimp, including a box, is in the enclosure. The chimp uses the box to get the bananas.
B. A lab-raised chimp is placed in the enclosure with bananas out of direct reach. None of the objects in the enclosure is familiar to the chimp. The chimp uses a bamboo shaft to reach the bananas.
C. The wild chimp described in the excerpt is placed in an enclosure with bananas out of direct reach. An assortment of objects, including the box the chimp used before, is in the enclosure. The chimp uses the box to get the bananas.
D. The wild chimp described in the excerpt is placed in an enclosure with bananas out of direct reach. The chimp uses an object that can serve as a platform to get the
E. A lab-raised chimp who has watched the banana problem being solved as described in passage 1 above is placed in an enclosure with bananas out of direct reach. An assortment of objects, including the box the chimp used before, is in the enclosure. The chimp uses the box to get the bananas.

**Question 3:** Köhler maintains that some animals are able to plan actions. When the bananas were mounted higher, the chimps again exhibited seemingly sudden insight, i.e. they utilized the objects around them to reach the bananas. Consider whether this observation (i) provides additional support for his argument; (ii) serves as evidence against his argument; or (iii) neither strengthens nor weakens the argument. In a few sentences, explain your choice of one of these alternatives.

**Passage 2:** Similar animal studies were performed by Edward C. Tolman of the University of California at Berkeley in the 1940s. He would allow a rat to explore a T-shaped maze with the same reward of food at either end. In one variant, the left arm ended in a dark, narrow box, whereas the right arm terminated in a wide, white box. (Rats prefer dark, narrow boxes.) On another day the rat was taken to a different room, placed in a dark, narrow box and electroshocked. On a subsequent day the rat was returned to the original maze. The rat moved directly to the right end of the box and its wide white box.

**Question 4:** Which studies, Köhler's chimp studies or Tolman's rat studies, best support the hypothesis that some animals are able to plan actions in advance? Explain your choice in a few sentences.

**Passage 3:** Flexibility in the use of innate alternatives may also be evidence for simple thinking. Two groups of ground-nesting birds, killdeers and plovers, have a variety of distraction ruses that are used to lure potential predators away from their eggs. Each display begins with the bird leaving the nest and moving inconspicuously to a location well away from its eggs. The set of possible performances ranges from simply calling from a highly visible spot to the complex feigning of a broken wing. There is even a highly realistic rodent-imitation ploy in which the bird scoots through the underbrush rustling provocatively and uttering mouselike squeaks. Each species also has a separate "startle" display designed to keep harmless animals such as deer from stumbling into the nest. Anecdotal reports have suggested that the decision to leave the nest to perform a display, as well as which display to employ, are suited to the degree of predator threat. A fox heading directly toward the nest, for example, is more likely to get the high-intensity broken-wing performance. Carolyn Ristau of Columbia University put this reported ability to gauge threats to a test by having distinctively dressed humans walk in straight lines near plover nests. Some were told to scan the ground carefully, apparently searching for nests, whereas others were instructed to pay no attention to the ground. As time went on, the plovers began to discriminate between
the potential hunters and the seemingly harmless humans: they did not even bother to leave the nest for the latter group but performed elaborate distraction displays for the former. Some degree of understanding seems evident in this ability to judge which innate response to select.

**Question 5.** Use the following information to devise a controlled experiment which measures reasoning by a bird. In 1-2 paragraphs, describe the experiment. Using the table below, describe how you will measure the bird's response, and the specific characteristics of your selected variable(s). Provide hypotheses for your experiment. Please explain one outcome that would support Ristau's claim, including potential confounding variables.

<table>
<thead>
<tr>
<th>Variable List for Potential Human Predator Conditions</th>
<th>Range of the variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>Short to tall</td>
</tr>
<tr>
<td>Proximity to bird nest</td>
<td>Near to far</td>
</tr>
<tr>
<td>Frequency of scanning behavior</td>
<td>Seldom to frequent</td>
</tr>
</tbody>
</table>

The suggested time for working on this question is 10-15 minutes.

**Passage 4:** Two well-studied cases of unusual foraging behavior in birds also suggest an ability to plan. One involves green herons - birds that capture fish using several (presumably inborn) approaches. In addition, occasionally herons have been observed bait fishing. They toss a morsel of food or a small twig into the water, and when a curious fish rises to investigate, the bird grabs it. Bait fishing has been observed in a few widely scattered spots in the U.S. and in a park in Japan. It appears on its own, seems (except once in Japan) not to spread to other birds and then vanishes. Given the high success rate of the technique, and yet the rarity of its use, bait fishing is probably not genetically programmed: most likely the trick has been independently invented by many different herons.

**Question 6.** What might counter the suggestion that bait fishing demonstrates reasoning by herons?

A. Given that bait fishing occurs only occasionally, the observation may not bear on innate reasoning ability of herons in general.
B. Perhaps too few observations have been made to conclude that bait fishing is unusual.
C. A repeated behavior caused by an accidental discovery could explain the bait fishing behavior.
D. Without an understanding of heron communication, any inference that bait fishing is the outcome of planning is speculative.
E. All of the above

**Question 7.** In a few sentences, please explain your answer to #6.
Part B: The following passages and questions were adapted from an article in the November, 2007 issue of Scientific American, "Cell Defenses and the Sunshine Vitamin," by Luz E. Tavera-Mendoza and John H. White.

**Passage 5:** It was called the sunshine cure, and in the early 20th century, before the era of antibiotics, it was the only effective therapy for tuberculosis known. No one knew why it worked, just that TB patients sent to rest in sunny locales were often restored to health. ...

The association [between low levels of exposure to sunlight and high incidence of] rickets [a disease leading to deformed and brittle bones]... steered most vitamin D research for the next 50 years toward understanding the molecule's role in bone building and how it acts in the kidneys, intestines and the skeleton itself to help control the flow of calcium into and out of bones from the bloodstream. In the past quarter century, however, studies of vitamin D function have broadened, revealing that the so-called sunshine vitamin does far more than build bones. Extensive evidence now shows that D has potent anticancer actions and also serves as an important regulator of immune system responses. Moreover, many of D's newly recognized benefits are maximized when it is present in the bloodstream at levels considerably higher than those found in many populations. These findings, together with epidemiological data linking low vitamin D levels to disease, support the possibility that widespread vitamin D deficiency is contributing to a number of serious illnesses [including cardiovascular diseases, cancers and autoimmune diseases such as multiple sclerosis].

**Question 8.** The passage refers to studies that indicate that vitamin D deficiencies related to low exposure to sunlight have a connection with incidence of tuberculosis, rickets, cancer, and immune system disorders such as multiple sclerosis. Which of the following reports describes the most scientifically rigorous connection of vitamin D to health?

A. A victim of multiple sclerosis was found to have a low level of vitamin D derivatives in her bloodstream.
B. Interviews with tuberculosis victims reveal that many of them spent little time in outdoor activities in their youth.
C. A group of women suffering from osteoporosis, a disease that leads to broken bones, had lower levels of vitamin D in their bloodstream than an otherwise comparable group of women who did not suffer loss of bone matter.
D. A higher rate of incidence of autoimmune diseases is observed at high latitudes, where the intensity of sunlight is lowest.
E. The population in America that has ancestry in northern Europe has generally lighter skin coloring than those who have ancestry in southern Europe.

**Passage 6:** The term "vitamin D" generally refers collectively to the two very similar molecules that come from each of those sources. Vitamin D3 ...is created by skin cells called
keratinocytes from a breakdown product of cholesterol, 7-dehydrocholesterol, in response to ultraviolet light. Vitamin D2, ... has slight structural differences that distinguish it from D3. Neither version has any biological activity in the body, however. First, either molecule must be modified by a series of related enzymes in a process called hydroxylation, to generate 25-hydroxyvitamin D (25D).

That conversion takes place primarily in the liver, but various cell types within the skin are also capable of performing the transformation locally. The 25D made by the liver is nonetheless the major form of vitamin D circulating in the bloodstream. When it is needed in the body, a final conversion to the biologically active form is required. [Specifically], 25D is further hydroxylated and becomes [the effective substance] 1,25-dihydroxyvitamin D (called 1,25D). The enzyme that performs this task, 1-alpha-hydroxylase, was first discovered in the kidney, and renal processing is responsible for generating much of the body's circulating 1,25D supply.

**Question 9.** Imagine a genetic disorder in which the hydroxylation enzymes are ineffective. Then the active 1,25D form of vitamin D is not produced. What conclusion is justified?

A. Exposure to sun is likely to relieve TB in patients with this genetic disorder.
B. Oral supplements of vitamin D should be provided to those with this disorder.
C. A blood test for 25D would show subnormal levels of this molecule
D. Administering 1,25D could be an effective therapy for this disorder.
E. None of the above

**Passage 7:** [M]apping studies of DNA ... revealed "Vitamin D Response Elements" [which form a Vitamin D-protein complex] lying close to two genes that encode antimicrobial peptides called cathelicidin and defensin beta 2. These small proteins act as natural antibiotics against a wide spectrum of bacteria, viruses and fungi. Fun a number of different cell types ... the rise in cathelicidin production [associated with the presence of vitamin D] was dramatic. [I]mmune cells treated with 25D, when exposed to pathogenic bacteria, released factors – presumably cathelicidin – that killed off the bacteria.

Philip Liu and Robert Modlin of the University of California, Los Angeles, and their collaborators ... induced the immune cells to start producing cathelicidin and to demonstrate antimicrobial activity against a variety of bacteria, including one that is perhaps the most intriguing: *Mycobacterium tuberculosis*. Thus, for the first time, the group revealed a plausible basis for the mysterious efficacy of the tuberculosis sunshine cure: the sun-soaked convalescents' vitamin D boost could have provided their immune cells with the raw material needed to generate a natural antibiotic that fought off the TB bacteria.

**Question 10.** If the conclusion that Vitamin D stimulates immune cells to combat TB bacteria is to be sound, what condition(s) must be met?
A. Vitamin D must be present in the cell-bacteria systems studied by Liu and Modlin.
B. Immune cells must interact directly with vitamin D.
C. Immune cells exposed to TB bacilli must transform vitamin D to its active form 1,25D.
D. Immune cells exposed to TB bacilli must not be able to combat the bacilli without vitamin D.
E. All of the above

Passage 8: Ultraviolet light penetrates the atmosphere more directly in the tropics than in more temperate regions of the planet, which receive substantial amounts of it only during the summer.

Because most people obtain vitamin D mainly through ultraviolet light exposure, circulating 25D levels in populations generally diminish with increasing latitude, although variations at a given latitude do arise because of differing ethnicities and diets, as well as variations in local climate and elevation. Consistent with vitamin D's observed gene-regulatory activities, a clear association is seen between increasing latitude and increased risk of several illnesses, most conspicuously autoimmune diseases such as MS.

Question 11. In any study intended to establish an association between latitude and incidence of MS, confounding factors can weaken the inference that exposure to sun is protective against MS. Choose any one of the list of confounding factors below and explain in a few sentences why that factor presents complications.

A. Diet: some subgroups of the population may obtain vitamin D from sources other than exposure to sun.
B. Gender: MS rates may be different for males and females
C. Season: vitamin D may be stored in fat so to maintain blood levels in dark weather.
D. Mobility: some people may live in more than one zone in the course of a year.
E. Latency: disorders may appear long after the deficiency has compromised the immune system.

Question 12 refers to the graphic below:

More detailed studies reveal complications not embraced in this simple generalization. Some such data are presented in the maps below.

<table>
<thead>
<tr>
<th>New MS cases each year per 100,000 in Norway</th>
<th>MS cases per 100,000 in Newfoundland</th>
<th>MS cases per 100,000 in Australia</th>
</tr>
</thead>
</table>

**Question 12**: Which statement is not justified by the data in the maps?

A. Australian data are well accounted for by the hypothesis that MS is correlated with latitude.
B. The data for Norway on the new cases occurring per year are roughly consistent with the data for Australia on the total of MS cases in a year.
C. Norwegian data suggest that proximity to ocean waters is strongly associated with low MS rates.
D. Newfoundland data suggest that proximity to the St John’s urban center is associated with low MS rates.
E. The narrow range of latitudes spanned by Newfoundland makes the data less relevant to the hypothesis.

http://www.direct-ms.org/sites/default/files/AlbertaDisadvantage.pdf
Passage 9: Studying 79 pairs of identical twins, scientists at the University of Southern California found an inverse relation between increased sun exposure during childhood and a lifetime risk of developing MS. In a comparison of MS rates between identical twins, the group of twins who spent more time outdoors as children had as much as 57 percent lower risk of developing the condition.

Question 13. Why is the experience of identical twins considered so important?

A. They share a common genetic makeup.
B. They are likely to have more accurate recollections of childhood behavior.
C. They share a common cultural background.
D. They are in the same age group.
E. They are generally of the same height, weight, and state of health.

The following two passages and questions were also adapted taken from "Cell Defenses and the Sunshine Vitamin."

Passage 10: In addition to the many studies correlating sun exposure with disease incidence, recent investigations have made similar connections between disease risk and levels of circulating 25D concentrations in blood serum. An enormous survey by researchers at the Harvard School of Public Health looked at the stored serum samples of some seven million U.S. Army and Navy personnel as well as their health records to see which individuals had developed MS between 1992 and 2004. The researchers found a significantly lower risk of eventual development of the disease in the group with high serum 25D levels at the time the sample was taken. Soldiers [and sailors] with serum 25D concentrations above 40 nanograms per milliliter had a 62 percent lower risk than the soldiers whose concentrations were 25 ng/ml or below.

Question 14: The following statements suggest that the study of the 25D level in serum is more convincing than the previous studies of seasonal variations in MS and latitude-incidence correlations. Which statement cannot be inferred from the passage above?

A. The number of samples is very large.
B. The difference of 62% lower risk in the group with high serum levels of 25D, compared with the group with low serum levels, is very unlikely to be due to random variation.
C. The study obeys the principle that those conducting tests on the serum should not know the medical history of the sample donor.
D. The sample population was young and healthy compared with the general population when samples were collected.
E. The study was conducted by researchers at a distinguished university.
Passage 11: Measuring circulating levels of 25D is the usual method of gauging vitamin D availability in the body. Generally agreed on health standards, based largely on bone-forming needs, hold circulating 25D levels of 30 to 45 ng/ml to be minimally sufficient. Serum vitamin D concentrations below between 21 and 29 ng/ml are considered insufficient and often accompanied by decreased bone density. Some symptoms of rickets may appear when concentrations fall below 20 ng/ml, and the risk of colon cancer rises.

Such low concentrations are unfortunately all too common, especially in winter. In February and March of 2005, for example, a survey of 420 healthy females across northern Europe – in Denmark (Copenhagen: 55° latitude), Finland (Helsinki: 60°), Ireland (Cork: 52°) and Poland (Warsaw: 52°) – found that 92 percent of adolescent girls in these countries had 25D levels less than 20 ng/ml and that 37 percent of the girls were severely deficient, with 25D levels of less than 10 ng/ml. Among the older women tested, 37 percent were found to be vitamin D deficient and 17 percent were severely deficient....

Vitamin D supplements could address the high prevalence of vitamin D deficiency in temperate zones, but how much people should take is still a subject of debate.

Question 15. Recommendations on the amount of vitamin D supplements required vary widely. The following issues may emerge in the debate on desirable amounts of vitamin D supplements. Choose any two of the following statements (all are true). Which of the two issues seems to you the more important public health issue? Explain why in a few sentences.

A. High levels of vitamin D dosage may be beneficial, but very high doses can be toxic.
B. Administered doses of vitamin D do not correlate perfectly with serum levels of 25D.
C. Vitamin D interacts with other common medications.
D. Lifestyle, skin color, diet, age, and other influences affect the amount of supplement required.
E. Vitamin D is available over the counter.

Part C: Questions 16 and 17 refer to the graph below, taken from a 1993 article in Scientific American, "How the Milky Way Formed," by Sidney van den Bergh and James E. Hesser.

Consider the graph below: the data (triangles and circles) refer to observed properties of two star clusters. The lines are generated from a mathematical model of brightness (luminosity) and temperature. The model relates these quantities to one another and the age of the stars in the cluster.
Question 16. Which statement characterizes the relation between luminosity (L) and color index (C) displayed in the graph above?

A. The L-C graph is linear (L = mC+b), or the relationship between luminosity and color is constant.
B. The L-C relations for clusters NGC-288 and NGC 362 are similar in form.
C. The theoretical forms (solid lines) are in serious disagreement with the experimental values (circles and triangles).
D. For a given value of luminosity, more than one value of color index can be read from the graph.
E. For a given value of color index only one value of luminosity can be read from the graph.

Question 17. According to the graph, star temperature

A. Has an inverse relationship with color index
B. Has an inverse relationship with luminosity
C. Has a similar relationship to color and luminosity for both NGC 288 and NGC 362
D. Has a non-linear relationship with visual magnitude.
E. All of the above
Part D: The following two excerpts were taken from a 2007 article in the *Los Angeles Times*, "Numbers Can Lie," by Andreas von Bubnoff.

**Passage 12:** In epidemiological studies (also called observational studies), scientists observe what’s going on; they don’t try to change it. From what they observe, they reach conclusions — for example, about the risk of developing a certain disease from being exposed to something in the environment, a lifestyle, or a health intervention.

There are different ways to do this. Cohort studies follow a healthy group of people (with intakes of, say, coffee) over time and look at who gets a disease. These are considered the strongest type of epidemiological study.

Case-control or retrospective studies examine people with and without a certain disease and compare their prior life — for how much coffee they drank, for example — and see if people who got the disease drank more coffee in their past than those who didn’t. Cross-sectional studies compare people’s present lifestyle (how much coffee they drink now) with their present health status.

Epidemiological studies have several advantages: they are relatively inexpensive, and they can ethically be done for exposures to factors such as alcohol that are considered harmful, because the people under study chose their exposure themselves. [However, none of these types of studies qualify as randomized and controlled experimental investigations.]

**Question 18.** Design a randomized and controlled experimental study to evaluate the effect of coffee drinking on the risk of developing pancreatic cancer. Include the essential characteristics of an experiment. List and describe four factors that need to be considered so to obtain valid results from such a study. In your study design, identify three important variables which should be measured and three confounding factors that must be considered.

The suggested time for working on this question is 10-15 minutes.

**Passage 13:** In a provocative 2005 paper, epidemiologist John Ioannidis examined the six most frequently cited epidemiological studies published from leading medical journals between 1990 and 2003. He found that in four of the six studies, the findings were later overturned by clinical trials. Vitamin E didn’t protect the heart in men or women. Hormone therapy didn’t protect the heart in women. Nitric oxide inhalation didn’t help patients with respiratory distress syndrome. Another finding turned out later to be exaggerated: taking flavonoids reduced coronary artery disease only by 20%, not by 68%, as originally reported.

The only finding that stood the test of time was a small study that reported that a chemical called all-trans retinoic acid was effective in treating acute promyelocytic leukemia.

The studies that overturned each of these epidemiological findings, Ioannidis says, "caused major waves of surprise when they first appeared, because everybody had believed the observational studies. And then the randomized trials found something completely different."

"Randomized studies often contradict one another, as do observational studies," says Harvard's Stampfer, who is an author on both of the frequently cited vitamin E and hormone replacement studies that Ioannidis says
were later refuted. Instead, Stampfer says, the two types of studies often test different things. "It's not an issue here that observational studies got it wrong and randomized trials got it right," he says, referring to the hormone replacement studies. "My view is that [both] were right and they were addressing different questions."

**Question 19. Which of the following might explain the conflict between epidemiological and clinical studies reported in the Ioannidis paper?**

A. The epidemiological findings mistook random relationships for significant effects  
B. The clinical trials mistook random relationships for significant effects  
C. Different populations were studied in the two methods  
D. Different questions were studied  
E. All of the above

**Rubric Question #3**

<table>
<thead>
<tr>
<th></th>
<th>4 = Highly Competent</th>
<th>3</th>
<th>2</th>
<th>1 = Not Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>General (unmapped to specific outcomes)</td>
<td>Clear and correct English; addresses question(s)</td>
<td>Multiple errors in English usage, mechanics, and grammar; fails to address question(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Understand that, while scientific statements are in principle tentative, criteria exist by which they can be judged, including consistency with the body of scientific theory, method, and knowledge.</td>
<td>Respondent uses pertinent data from both studies to defend their position. Avoids unsupported assertions.</td>
<td>Respondent does not use pertinent data from both studies to defend their position; Makes several unsupported assertions.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Display a grasp of experimental design, including the notion of the control and the idea of statistical significance.</td>
<td>Recognizes that the new task is more difficult than the first; i.e. the 2nd experiment is not a simple extrapolation of the 1st. Identifies limitations of the study.</td>
<td>Does not recognize that the new task is more difficult than the first. Does not identify limitations of the study.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Acknowledge the possibility of alternative accounts of events and properties and judge their relative plausibility by standard criteria.</td>
<td>Recognizes that the 2nd study does not prove that chimps plan actions in advance but rather supports the hypothesis. Recognizes that other experimental factors (e.g. learned behavior, instinct, etc.) must be</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

36
learned behavior, instinct, etc.) must be considered.

Model answer: The observation provides additional support for Kohler’s hypothesis. The chimps solve a similar problem but one which is considerably more difficult. (Instead for reaching bananas by standing on a singly box, the chimps must use multiple objects, staked on top of one another, to reach the bananas.) To solve the new problem, the chimps must apply a novel, creative solution. Nevertheless, additional studies are required to account for other possibilities for the observed behavior (e.g. learned behavior, instinct, etc.) Thus, additional studies which vary the environment of the chimps and variables other than height are required.

Rubric Question #4

<table>
<thead>
<tr>
<th>General (unmapped to specific outcomes)</th>
<th>4 = Highly Competent</th>
<th>3</th>
<th>2</th>
<th>1 = Not Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear and correct English; addresses question(s)</td>
<td>Multiple errors in English usage, mechanics, and grammar; fails to address question(s)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carefully compares, evaluates and judges the two studies. Makes a clear and logically valid concluding statement.</td>
<td>Fails to compare studies. Does not make a clear and logically valid concluding statement.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Refers to controls and the role of statistical judgment in the design of the studies.</td>
<td>Fails to address studies' designs.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addresses both studies and incorporates information about each. Identifies relevant differences between the two.</td>
<td>Does not compare studies.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addresses the issue of interpretation of animal behavior.</td>
<td>Overlooks possibility of error.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Model answer: While both studies support the hypothesis that some animals are better able to conceive of a plan to accomplish a goal, Tolman’s experiments are more convincing. Tolman’s studies account for several of the variables which plague Kohler’s experiments. Specifically, Tolman better accounts for the environment. For example, the rats are placed in a maze which has no differential reinforcement, whereas the apes are reinforced by the bananas. Also, the rats are placed in an environment which accounts for their instinctual reactions, i.e. the study controls for their preference for small, dark places, whereas one cannot eliminate instinctual reactions for the chimps. Finally, since Tolman’s rats are only introduced to the maze once before and then once during the experiment, this minimizes the influences of a learned behavior. Tolman’s study is not without flaws, though. For instance, one could argue that the rats are conditioned to avoid the preferred black box.

Rubric Question #5

<table>
<thead>
<tr>
<th>4 = Highly Competent</th>
<th>3</th>
<th>2</th>
<th>1 = Not Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>General (unmapped to specific outcomes)</td>
<td>Clear and correct English; addresses question(s)</td>
<td>Multiple errors in English usage, mechanics, and grammar; fails to address question(s)</td>
<td></td>
</tr>
<tr>
<td>1. Understand that, while scientific statements are in principle tentative, criteria exist by which they can be judged, including consistency with the body of scientific theory, method, and knowledge.</td>
<td>Acknowledges that the experiment cannot serve as proof of reasoning capacity in birds.</td>
<td>Accepts that the experiment’s findings indicating any bird reaction is a result of reasoning.</td>
<td></td>
</tr>
<tr>
<td>2. Display a grasp of experimental design, including the notion of the control and the idea of statistical significance.</td>
<td>Addresses all variables including the new height variable. Makes a clear hypothesis; specifies bird’s reaction and outcomes.</td>
<td>Omits or provides only vague characterization of variables, hypotheses, descriptions of birds’ responses and outcomes.</td>
<td></td>
</tr>
<tr>
<td>5. Identify sources of error in scientific investigation, including errors of measurement and ambiguity of judgment.</td>
<td>Identifies confounding factors (and describes remedies).</td>
<td>No consideration of confounding factors, or gives illogical or implausible confounding factors with unsound remedies.</td>
<td></td>
</tr>
</tbody>
</table>

Model answer: **Hypothesis**: Bird’s calling response is hypothesized to increase as the predator comes closer to the next and increases his scanning behavior. Height is not assumed to affect the bird’s calling response, but height will be tested. Every variable will be combined with every other variable: **Predator**
height: (Short (e.g., about 5ft) medium (about 5’6’’), tall (about 6’)), Proximity to bird nest: (near (e.g., about 10 ft), moderate (about 20 ft), far (about 30 ft)), Frequency of scanning behavior (seldom (e.g., about once per 30 seconds), moderate (about three times per 30 seconds), frequent (about six times per 30 seconds)). (This description would consist of 3 x 3 x 3 conditions, or 27 total conditions. Other model answers may include fewer variables and levels, for example, only 2 levels of nest proximity and 2 levels of scanning frequency, for a total of 4 conditions). The bird’s calling response (or departure from nest, or feigning a broken wing response, or squeak response, e.g.) will be measured by videotaping bird and timing the latency to begin a calling response from the time the predator begins walking near the nest. Measurements will be taken for (e.g.) a two-minute time period. Bird’s responses will be potentially reduced to a categorical response if bird responses are sparse and reduced precision results in no loss of information. Supported outcome: Bird’s response increased in frequency (or intensity, etc.) as scanning frequency and proximity increased. Potential confounds: the physical appearance of the human predator may be a confound if not held constant across predator height levels or differences in predator walking style (i.e., some predators walk more heavily than others), predator gender, etc.

Rubric Question 7

<table>
<thead>
<tr>
<th></th>
<th>4 = Highly Competent</th>
<th>3</th>
<th>2</th>
<th>1 = Not Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>General (unmapped to specific outcomes)</td>
<td>Clear and correct English; addresses question(s)</td>
<td>Multiple errors in English usage, mechanics, and grammar; fails to address question(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Understand that, while scientific statements are in principle tentative, criteria exist by which they can be judged, including consistency with the body of scientific theory, method, and knowledge.</td>
<td>Explanation goes beyond a paraphrase of the choice. Refers to passage.</td>
<td>Makes statements with unwarranted certainty. Explanation is unrelated to the passage and response options.</td>
<td>Fails to address how or why the choice(s) indicate(s) an alternative explanation of behavior.</td>
<td></td>
</tr>
<tr>
<td>4. Acknowledge the possibility of alternative accounts of events and properties and judge their relative plausibility by standard criteria.</td>
<td>Addresses the question; if choice is E, explains why A-D are all plausible explanations of the behavior.</td>
<td>Fails to address how or why the choice(s) indicate(s) an alternative explanation of the behavior.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Identify sources of error in scientific investigation, including errors of measurement and ambiguity of judgment.</td>
<td>Refers to the rarity of the observed behavior, danger of attributing motive and intent to bird behavior.</td>
<td></td>
<td>Overlooks these sources of error.</td>
<td></td>
</tr>
</tbody>
</table>
6. Recognize unsound conclusions.

**Explanation is logical. If choices are A-D, explanation shows how the objection counters the general assertion of heron reasoning. If choice is E, all of these explanations should be included.**

**If choice is A-D, rationale is absent or fails to explain the basis for the objection. If choice is E, fails to address all of these objections.**

**Model Answer:** Perhaps A or B. All seem to be serious objections, but they seem to bear on the difficulty of definition of the subject in question. A very few observations make a slender basis for grand inferences, regardless of whether bait-fishing is rare or widespread. The nature of reasoning is not very clearly defined – isn’t putting a discovery to later use learning and reasoning? As for communication, isn’t teaching possible by wordless example? “Planning” adds another dimension to the confusion. A model answer would be a serious confrontation with the issues regardless of the actual choice A-E.

**Rubric Question #11**

<table>
<thead>
<tr>
<th></th>
<th>4 = Highly Competent</th>
<th>3</th>
<th>2</th>
<th>1 = Not Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General (unmapped to specific outcomes)</strong></td>
<td>Clear and correct English; addresses question(s)</td>
<td>Multiple errors in English usage, mechanics, and grammar; fails to address question(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2. Display a grasp of experimental design, including the notion of the control and the idea of statistical significance.</strong></td>
<td>Reply reflects recognition of the need to control a variable.</td>
<td>No mention of control.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Quality of Explanation</strong></td>
<td>Explanation goes beyond a paraphrase of the choice.</td>
<td>Conscious refusal to explain choice.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>5. Identify sources of error in scientific investigation, including errors of measurement and ambiguity of judgment.</strong></td>
<td>Addresses question; Reply makes clear how the proposed confounding factor can cause variation in MS rates independent of latitude or can complicate measurements.</td>
<td>No rationale for choice.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Model Answer:

**If A:** Diet changes with latitude because of differences in locally available foods or other changes in culture. If populations in high latitudes consume foods rich in vitamin D, this would weaken the observed latitude – D deficiency correlation.

**If B:** Samples from the populations must be sex-specific or else introduce sex as a variable; if the high latitude population sample is weighted by the less susceptible sex, this would weaken the observed correlation.

**If C:** Exposure to sun during the long summer at northern latitudes may be protective through the dark winter; the total annual exposure may not be so important as some minimum dose.

**If D:** If the population is mobile a reliable estimate of exposure to sunlight cannot be inferred by subject location when interviewed.

**If E:** Current latitude may not bear on the degree of exposure at a critical time in the course of the illness.

---

**Rubric Question #15**

<table>
<thead>
<tr>
<th>4 = Highly Competent</th>
<th>3</th>
<th>2</th>
<th>1 = Not Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>General (unmapped to specific outcomes)</td>
<td>Clear and correct English; addresses question(s)</td>
<td>Multiple errors in English usage, mechanics, and grammar; fails to address question(s)</td>
<td></td>
</tr>
<tr>
<td>1. Understand that, while scientific statements are in principle tentative, criteria exist by which they can be judged, including consistency with the body of scientific theory, method, and knowledge.</td>
<td>Addresses question: draws plausible inferences on relative impact of each source of difficulty in establishing dosage.</td>
<td>Fails to choose and compare two issues. Relies on assertions.</td>
<td></td>
</tr>
<tr>
<td>5. Identify sources of error in scientific investigation, including errors of measurement and ambiguity of judgment.</td>
<td>Explains why the two chosen are problematic for public health.</td>
<td>Does not explain why the two chosen are problematic for public health.</td>
<td></td>
</tr>
</tbody>
</table>
Model Answer: Since there are $5 \times 4 = 20$ combinations and two possible judgments for each pair, we will show only a few characteristic responses: A seems most urgent since there are deaths to consider. The public health campaign should emphasize than “more = better” does not apply to vitamin D. B is part of the usual uncertainty typical of dosage judgments; its significance depends on the seriousness of any discrepancy. C is important to the degree interaction is a large effect, and is a matter for evaluation and if need be container labeling. D (like A-C) requires some sort of understanding of the range of safe/beneficial doses. E becomes an issue only if customers are encouraged to take huge doses. The key public health matters seem to me to be: to establish a valid dosage range and to establish the degree and kind of drug interaction with vitamin D. A model reply need not come to any specific conclusion, but should make a point and support it.
**Rubric Question #18**

<table>
<thead>
<tr>
<th></th>
<th>4 = Highly Competent</th>
<th>3</th>
<th>2</th>
<th>1 = Not Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>General (unmapped to specific</td>
<td>Clear and correct</td>
<td>Multiple errors in English usage,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>outcomes)</td>
<td>English; addresses</td>
<td>mechanics, and grammar; fails to address question(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>question(s)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Understand that, while</td>
<td>Notes need for data</td>
<td>Notes need for data analysis comparing cancer rates.</td>
<td>Does not discuss measurement of cancer rates in two groups.</td>
<td></td>
</tr>
<tr>
<td>scientific statements are</td>
<td>analysis comparing</td>
<td>Notes need for assessment on multiple occasions.</td>
<td>Does not address statistical comparison of rates.</td>
<td></td>
</tr>
<tr>
<td>in principle tentative,</td>
<td>cancer rates.</td>
<td>Difference needs to reach some pre-set criterion.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>criteria exist by which they</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>can be judged, including</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>consistency with the body of</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scientific theory, method, and</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>knowledge.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Display a grasp of</td>
<td>Random assignment</td>
<td>No mention of random assignment or need for representative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>experimental design, including</td>
<td>with notion of</td>
<td>sample.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the notion of the control and</td>
<td>control group,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the idea of statistical</td>
<td>Representative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>significance.</td>
<td>sample, Profiling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>the population.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Identify sources of error</td>
<td>Cites three</td>
<td>Does not consider possible confounding factors.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in scientific investigation,</td>
<td>confounding factors.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>including errors of</td>
<td>Notes possible</td>
<td>Does not mention possible relationship link between some agent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>measurement and ambiguity of</td>
<td>alternative causes</td>
<td>and cancer rate.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>judgment.</td>
<td>of cancer.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Considers coffee</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>might not be driver</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>of pancreatic cancer.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Model Answer:** Two groups of individuals need to be available for the subject pool — coffee drinkers and non-coffee drinkers. The sample should be representative of all adults and of sufficient size to detect any effects that might be present but not so large that minor effects reach the preset criterion for statistical significance. Subjects should be randomly selected from each group. A health assessment including family history should be obtained from each Subject. For the coffee drinkers, daily logs of the amount of coffee and type of coffee (regular vs. decaf) consumed must be maintained. For all Subjects, records of food consumed, smoking behavior, alcohol consumption, and other potential carcinogenic factors should be obtained. Annual (at least) medical examinations need be obtained for all Subjects and the incidents of pancreatic cancer recorded. The study should be of sufficient duration to allow for the development of cancer. Factors to consider should also include age, race, and gender of Subjects. Subjects in the non-coffee drinking group must immediately report the consumption of coffee.
### Appendix 4: Scoring on Individual Questions - Longitudinal Assessment

**Table 1**

<table>
<thead>
<tr>
<th>LO1</th>
<th>LO2</th>
<th>LO3</th>
<th>LO4</th>
<th>LO5</th>
<th>LO6</th>
<th>Question</th>
<th>Question Type</th>
<th>As 1st Years</th>
<th>As 4th Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Q1</td>
<td>Multiple-Choice*</td>
<td>1.59 40%</td>
<td>2.97 74%</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>Q2</td>
<td>Multiple-Choice</td>
<td>2.47 62%</td>
<td>2.51 63%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>Q3</td>
<td>Short Answer</td>
<td>3.43 57%</td>
<td>3.83 64%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Q4</td>
<td>Short Answer†</td>
<td>3.27 55%</td>
<td>3.61 60%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>Q5</td>
<td>Experimental Design</td>
<td>8.56 57%</td>
<td>9.5 63%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Q6</td>
<td>Multiple-Choice</td>
<td>“set up” for Q7, not scored</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>Q7</td>
<td>Short Answer</td>
<td>4.09 68%</td>
<td>4.01 67%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td>X</td>
<td>Q8</td>
<td></td>
<td>Multiple-Choice</td>
<td>3.12 78%</td>
<td>3.2 75%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>Q9</td>
<td>Multiple-Choice</td>
<td>2.25 56%</td>
<td>2.51 63%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td>X</td>
<td>Q10</td>
<td>Multiple-Choice</td>
<td>0.49 12%</td>
<td>0.91 23%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>Q11</td>
<td>Short Answer*</td>
<td>3.56 59%</td>
<td>4.28 80%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td>Q12</td>
<td>Multiple-Choice</td>
<td>2.14 53%</td>
<td>2.63 66%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td>Q13</td>
<td>Multiple-Choice</td>
<td>3.56 89%</td>
<td>3.89 97%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>Q14</td>
<td>Multiple-Choice</td>
<td>2.63 66%</td>
<td>3.09 77%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td>Q15</td>
<td>Short Answer*</td>
<td>2.99 50%</td>
<td>3.76 63%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td>Q16</td>
<td>Multiple-Choice</td>
<td>2.74 69%</td>
<td>2.74 69%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td>Q17</td>
<td>Multiple-Choice</td>
<td>1.15 29%</td>
<td>0.57 13%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>Q18</td>
<td>Experimental Design</td>
<td>10.00 67%</td>
<td>10.47 70%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td>Q19</td>
<td>Multiple-Choice</td>
<td>2.19 55%</td>
<td>2.4 60%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: “LO” is Learning Outcome; "%" is the average percent of attainable points.

* p<.05  †p=09

**LO 1** - Understands that while scientific statements are in principle tentative, criteria exist by which their validity can be judged, including consistency with the body of scientific theory, method, and knowledge;

**LO 2** - Displays a grasp of experimental design, including the notion of the control and the idea of statistical significance of differences in outcome

**LO 3** - Infers valid information from quantitative data presented in graphical form

**LO 4** - Acknowledges the possibility of alternative accounts of events and properties and judges their relative plausibility

**LO 5** - Identifies sources of error in scientific investigation, including errors of measurement and ambiguity of judgment

**LO 6** - Detects unsound logic in scientific arguments
Figure 1: Mean Scores by Multiple Choice Question, First- vs. Fourth-Years

Figure 2: Mean Scores by Experimental Design Question, First- vs. Fourth-Years
Figure 3: Mean Scores by Short-Answer Question, First- vs. Fourth-Years

<table>
<thead>
<tr>
<th>Question</th>
<th>1ST</th>
<th>4TH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q3</td>
<td>3.59</td>
<td>3.83</td>
</tr>
<tr>
<td>Q4~</td>
<td>3.23</td>
<td>3.61</td>
</tr>
<tr>
<td>Q7</td>
<td>4.01</td>
<td>4.23</td>
</tr>
<tr>
<td>Q11*</td>
<td>3.61</td>
<td>4.28</td>
</tr>
<tr>
<td>Q15*</td>
<td>3.12</td>
<td>3.76</td>
</tr>
</tbody>
</table>

* p<.05
** p=.09
Appendix 5: 2013 Revised Scientific Reasoning Test and Associated Rubrics

UNIVERSITY OF VIRGINIA CORE COMPETENCY ASSESSMENT
Scientific Reasoning
2013

Instructions:

You may begin the test after you have read through these instructions. You will have 50 minutes to complete the assessment.

This instrument was developed by University faculty and administrators, with input from students, and is being administered for the sole purpose of improving undergraduate education. Your score will never be a part of your academic record. It will be used to compute composite scores for the University, your school, and perhaps your major.

The goal of this assessment is to help the University better understand how well our students can reason using scientific principles and methods. We want to assess how well the University is teaching science reasoning to all undergraduates, not just science majors.

The assessment contains five short answer questions and asks you to design two studies. Each question requires you to read a passage from a published article and then respond to the question posed. Please answer all questions. The suggested working time for short answer questions is no more than 5 minutes, the suggested working time for study designs is no more than 10 minutes.

For these written components, please keep in mind that we are assessing scientific reasoning skills, not writing ability, but faculty evaluators must be able to understand what you have written.

The assessment was designed to be taken in the order the questions appear. You may skip difficult questions though, and return to them as you wish.

Some of the passages have been edited for the purposes of designing this instrument. In cases where the passage has been edited, ellipses (...) or brackets [ ] denote removed or added text, respectively.

Although this assessment is not course-based or for a grade, please do your best work. Thank you for your participation!
The following passages and questions are based on a 1998 article in Scientific American, “Reasoning in Animals,” by James L. Gould and Carol Grant Gould.

Passage 1: The ability to think and plan is taken by many of us to be the hallmark of the human mind. Reason, which makes thinking possible, is often said to be uniquely human and thus sets us apart from the beasts. In the past two decades, however, this comfortable assumption of intellectual superiority has come under increasingly skeptical scrutiny. Most researchers now at least entertain the once heretical possibility that some animals can indeed think.

In an early, controversial study (1914), German psychologist Wolfgang Köhler was working at a primate research center on the Canary Islands. He presented novel problems to chimps he had caught in the wild. Often the pattern of solution suggested insight rather than trial and error. For instance, when Köhler first hung a bunch of bananas out of reach, the chimpanzee being observed made a few useless leaps, then went off to a corner and "sulked." But in time the chimp looked back at the bananas, then around the large outdoor enclosure at the various objects he had to play with, back to the bananas, back to one specific toy (a box), then ran directly to the box, dragged it under the fruit, climbed on top, leaped up and grabbed the prize.

Question 1: Köhler maintains that some animals are able to plan actions. When the bananas were mounted higher, the chimps again exhibited seemingly sudden insight, i.e., they utilized the objects around them to reach the bananas. Consider whether this observation (i) provides additional support for his argument; (ii) serves as evidence against his argument; or (iii) neither strengthens nor weakens the argument. In a few sentences, explain your choice of one of these alternatives.

(space provided for answer on test)
Passage 2: Similar animal studies were performed by Edward C. Tolman of the University of California at Berkeley in the 1940s. He would allow a rat to explore a T-shaped maze with the same reward of food at either end. In the maze, the left arm ended in a dark, narrow box, whereas the right arm terminated in a wide, white box. (Rats prefer dark, narrow boxes.) On another day the rat was taken to a different room, placed in a dark, narrow box and electroshocked. On a subsequent day the rat was returned to the original maze. The rat moved directly to the right end of the maze and its wide white box.

Question 2: Which studies, Köhler’s chimp studies or Tolman’s rat studies, best support the hypothesis that some animals are able to plan actions in advance? Explain your choice in a few sentences.

(space provided for answer on test)
Passage 3: *Flexibility in the use of innate alternatives may also be evidence for simple thinking. Two groups of ground-nesting birds, killdeers and plovers, have a variety of distraction ruses that are used to lure potential predators away from their eggs. Each display begins with the bird leaving the nest and moving inconspicuously to a location well away from its eggs. The set of possible performances ranges from simply calling from a highly visible spot to the complex feigning of a broken wing. There is even a highly realistic rodent-imitation ploy in which the bird scoots through the underbrush rustling provocatively and uttering mouse-like squeaks. Each species also has a separate "startle" display designed to keep harmless animals such as deer from stumbling into the nest. Anecdotal reports have suggested that the decision to leave the nest to perform a display, as well as which display to employ, are suited to the degree of predator threat. A fox heading directly toward the nest, for example, is more likely to get the high-intensity broken-wing performance. Carolyn Ristau of Columbia University put this reported ability to gauge threats to a test by having distinctively dressed humans walk in straight lines near plover nests. Some were told to scan the ground carefully, apparently searching for nests, whereas others were instructed to pay no attention to the ground. As time went on, the plovers began to discriminate between the potential hunters and the seemingly harmless humans: they did not even bother to leave the nest for the latter group but performed elaborate distraction displays for the former. Some degree of understanding seems evident in this ability to judge which innate response to select.*

**Question 3:** Use the following information to devise a controlled experiment that measures reasoning by a bird. In 1-2 paragraphs, describe the experiment. Using the table below, describe how you will measure the bird’s response and the specific characteristics of your selected variable(s). Provide hypotheses for your experiment. Please explain one outcome that would support Ristau’s claim, including potential confounding variables.

<table>
<thead>
<tr>
<th>Variable List for Potential Human Predator Conditions</th>
<th>Range of the variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>Short to tall</td>
</tr>
<tr>
<td>Proximity to bird nest</td>
<td>Near to far</td>
</tr>
<tr>
<td>Frequency of scanning behavior</td>
<td>Seldom to frequent</td>
</tr>
</tbody>
</table>

The suggested time for working on this question is 10 minutes.

(space provided for answer on test)
Passage 4: Two well-studied cases of unusual foraging behavior in birds also suggest an ability to plan. One involves green herons—birds that capture fish using several (presumably inborn) approaches. In addition, occasionally herons have been observed bait fishing. They toss a morsel of food or a small twig into the water, and when a curious fish rises to investigate, the bird grabs it. Bait fishing has been observed in a few widely scattered spots in the U.S. and in a park in Japan. It appears on its own, seems (except once in Japan) not to spread to other birds and then vanishes. Given the high success rate of the technique, and yet the rarity of its use, bait fishing is probably not genetically programmed: most likely the trick has been independently invented by many different herons.

What reasoning best counters the suggestion that bait fishing demonstrates planning by herons? (mark one)

A. Given that bait fishing occurs only occasionally, the observation may not bear on innate planning ability of herons in general.
B. Perhaps too few observations have been made to conclude that bait fishing is unusual.
C. An accidental behavior that leads to success could explain the bait fishing.
D. Without an understanding of heron communication, any inference that bait fishing is the outcome of planning is speculative.

**Question 4:** In a few sentences, please explain your choice.

(space provided for answer on test)
The following passages and questions were adapted from an article in the November 2007 issue of Scientific American, “Cell Defenses and the Sunshine Vitamin,” by Luz E. Tavera-Mendoza and John H. White.

**Passage 5:** As shown in the figure below, ultraviolet light penetrates the earth’s atmosphere more easily at the equator than at other latitudes.

![Vitamin D Winter Diagram](image)

Because most people obtain vitamin D mainly through ultraviolet light exposure, circulating vitamin D levels in populations generally diminish with increasing latitude, although variations within latitudes do arise because of differences in ethnicities and diets, local climate and elevation.

**Question 5.** The results of some studies have established an association between increasing latitude and increased risk of several illnesses such as multiple sclerosis (MS). Referring to the figure above, suggest two confounding factors that would weaken the inference that exposure to sun is protective against MS. For each confounding factor you propose, explain in a few sentences why that factor presents complications.

(space provided for answer on test)
The following passage and question were adapted from a February 2013 article in the British Journal of Sports Medicine, “The effects of Vitamin D3 supplementation on serum total 25(OH)D concentration and physical performance: a randomized dose-response study” by G.L. Close, J. Leckey, and M. Paterson.

**Passage 6:** Given the importance of vitamin D in human physiology and disease prevention, an alarming number of populations exhibit low vitamin D concentration, even in healthy, athletic populations. Vitamin D can be obtained from dietary sources, although it is primarily synthesized endogenously from solar ultraviolet B (UVB) radiation.

The biologically active form of vitamin D in the body is called “1,25-dihydroxyvitamin D3” or 25(OH)D. Reviewing the data on 25(OH)D concentrations of athletes and the physiological consequences of any deficiency is complicated by the ongoing debate as to what concentrations of total 25(OH)D constitute vitamin D deficiency and toxicity. Zittermann (2003) defines optimum 25(OH)D concentration as being between 100 and 250 nmol/l, whereas the US Institute of Medicine defines deficient 25(OH)D concentration as being less than 50 nmol/l and suggests that there are potential adverse events associated with levels greater than 125 nmol/l.

**Research Design:** Researchers completed a double-blind study i.e. researchers and participants do not know to which group each participant belongs to test the effects of various levels of vitamin D supplementation on 30 University club-level athletes’ blood 25(OH)D levels. Participants received either 20,000 or 40,000 IU vitamin D3 or a visually identical placebo (PLB) once a week for 12 weeks.

**Results of the experiment are presented in the graph.**

Change in serum total 25(OH)D concentration following suppression with PLB, 20,000 and 40,000 IU/week of vitamin D3

Change in serum total 25(OH)D concentration following supplementation with a placebo, 20,000 and 40,000 IU/week of vitamin D3. Line with * indicates significant difference from Pre in all groups at the corresponding time point and # indicates significant difference from the 40,000 IU group (p<0.05).
Question 6: An athletic coach believes the optimal serum level is 75nmol/l and wants one of her athletes to increase their 25(OH)D to an optimal level before an upcoming tournament. Using information presented in the graph, explain the dose of Vitamin D₃ (0 IU, 20,000 IU, or 40,000 IU) and length of supplementation, if any, that you would recommend to the athlete. Explain if your recommendation would change if the coach believed the optimal 25D serum level to be 100umol/l.

(space provided for answer on test)

Question 7: Recommendations on the amount of vitamin D supplements required vary widely. The following issues may emerge in the debate on desirable amounts of vitamin D supplements. Choose any two of the following statements (all are true). Compare and contrast the two statements you chose and explain which of the two issues seems to you the more important public health issue. Explain why in a few sentences.

A. High levels of vitamin D dosage may be beneficial, but very high doses can be toxic.
B. Administered doses of vitamin D do not correlate perfectly with serum levels of 25D.
C. Vitamin D interacts with other common medications.
D. Lifestyle, skin color, diet, age, and other influences affect the amount of supplement required.
E. Vitamin D is available over the counter.

(space provided for answer on test)
The following excerpt was taken from an article in the *Los Angeles Times*, "Numbers Can Lie," by Andreas von Bubnoff.

**Passage 7:** In epidemiological studies (also called observational studies), scientists observe what's going on; they don't try to change it. From what they observe, they reach conclusions — for example, about the risk of developing a certain disease from being exposed to something in the environment, a lifestyle, or a health intervention.

There are different ways to do this. Cohort studies follow a healthy group of people (with intakes of, say, coffee) over time and look at who gets a disease. These are considered the strongest type of epidemiological study.

Case-control or retrospective studies examine people with and without a certain disease and compare their prior life — for how much coffee they drank, for example — and see if people who got the disease drank more coffee in their past than those who didn't. Cross-sectional studies compare people's present lifestyle (how much coffee they drink now) with their present health status.

Epidemiological studies have several advantages: they are relatively inexpensive, and they can ethically be done for exposures to factors such as alcohol that are considered harmful, because the people under study chose their exposure themselves. [However, none of these types of studies qualify as randomized and controlled experimental investigations.]

**Question 8:** Design an observational study to evaluate the effect of coffee drinking on the risk of developing pancreatic cancer. Include the essential characteristics of a well-designed observational study. In your study design, list and describe four factors that need to be considered in order to obtain valid results from such a study, identify and describe three important variables that should be measured, and identify and describe three confounding factors that must be considered.

The suggested time for working on this question is 10 minutes.

(space provided for answer on test)

IF YOU ARE A FIRST-YEAR, PLEASE TELL US WHAT YOUR INTENDED MAJOR IS (OR IF YOU DON’T KNOW YET, PUT YOUR BEST GUESS):

MAJOR: _____________________________________________________

THANK YOU FOR COMPLETING THE 2013 SPRING ACADEMIC ASSESSMENT!
Rubrics for Short-Answer and Design Questions in the

2013 Test, with Possible Answers
Question #1  SCORE (0-6) ______________

<table>
<thead>
<tr>
<th>Scoring Guide</th>
<th>(0-1) Not Competent</th>
<th>(1-2) Minimally Competent</th>
<th>(3-4) Competent</th>
<th>(5-6) Highly Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Understand that, while scientific statements are in principle tentative, criteria exist by which they can be judged, including consistency with the body of scientific theory, method, and knowledge.</td>
<td>Respondent does not use pertinent information from both studies to defend their position; fails to compare studies. Makes unsupported assertion(s).</td>
<td></td>
<td></td>
<td>Respondent uses pertinent information from both studies to defend their position. Avoids unsupported assertion(s).</td>
</tr>
</tbody>
</table>

**Clarity of writing:** Competent responses are generally clearly written using correct English. Minimally competent or not competent writing is that which fails to address the question posed, or is unclear, unorganized or contains errors of English usage, mechanics, or grammar that impede the evaluator’s comprehension of the response.

Competent □ Minimally competent □ Not Competent □

**One example of a highly competent response:** The observation provides additional support for Kohler’s hypothesis. The chimps solve a similar problem but one that is considerably more difficult. (Instead of reaching bananas by standing on a single box, the chimps must use multiple objects, stacked on top of one another, to reach the bananas.) To solve the new problem, the chimps must apply a novel, creative solution. Nevertheless, additional studies are required to account for other possibilities for the observed behavior (e.g. learned behavior, instinct, etc.) Thus, additional studies should vary the environment of the chimps and take into account variables other than height.
Institutional Assessment and Studies

Scientific Reasoning 2013

Question #2  SCORE (0-6) _______________

<table>
<thead>
<tr>
<th>Scoring Guide</th>
<th>(0-1) Not Competent</th>
<th>(1-2) Minimally Competent</th>
<th>(3-4) Competent</th>
<th>(5-6) Highly Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Understand that, while scientific statements are in principle tentative,</td>
<td>Fails to compare studies.</td>
<td>Does not make a clear and logically valid</td>
<td>Carefully compares, evaluates or judges</td>
<td>Makes a clear and logically valid conclusions.</td>
</tr>
<tr>
<td>criteria exist by which they can be judged, including consistency with the</td>
<td></td>
<td>concluding statement.</td>
<td>the two studies.</td>
<td></td>
</tr>
<tr>
<td>body of scientific theory, method, and knowledge.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Clarity of writing:** Competent responses are generally clearly written using correct English. Minimally competent or not competent writing is that which fails to address the question posed, or is unclear, unorganized or contains errors of English usage, mechanics, or grammar that impede the evaluator’s comprehension of the response.

Competent □  Minimally competent □  Not Competent □

**One example of a highly competent response:** While both studies support the hypothesis that some animals are better able to conceive of a plan to accomplish a goal, Tolman’s experiments are more convincing. Tolman’s studies account for several of the variables that plague Kohler’s experiments. Specifically, Tolman better accounts for the environment. For example, the rats are placed in a maze with no differential reinforcement, whereas the apes are reinforced by the bananas. Also, the study controls for the rats’ instinct—their preference for small, dark places, whereas one cannot eliminate the role of instinct for the chimps. Finally, since Tolman’s rats are only introduced to the maze once before and then once during the experiment, this minimizes the influences of a learned behavior. Tolman’s study is not without flaws, though. For instance, one could argue that the rats are conditioned to avoid the preferred black box.
Question #3  SCORE (0-12) _______________

<table>
<thead>
<tr>
<th>Scoring Guide</th>
<th>(0-2) Not Competent</th>
<th>(3-5) Minimally Competent</th>
<th>(6-9) Competent</th>
<th>(10-12) Highly Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Display a grasp of experimental design, including the notion of the control.</td>
<td>Omits or provides only vague or general characterization of variables, hypotheses, descriptions of birds’ responses and outcomes.</td>
<td></td>
<td></td>
<td>Addresses all variables including the new height variable. Makes a clear hypothesis; specifies measurement of bird’s reaction and potential outcomes.</td>
</tr>
</tbody>
</table>

**Clarity of writing:** Competent responses are generally clearly written using correct English. Minimally competent or not competent writing is that which fails to address the question posed, or is unclear, unorganized or contains errors of English usage, mechanics, or grammar that impede the evaluator’s comprehension of the response.

Competent □ Minimally competent □ Not Competent □

**One example of a highly competent response:** Hypothesis: Bird’s calling response is hypothesized to increase as the predator comes closer to the nest and increases scanning behavior. Height is not assumed to affect the bird’s calling response, but height will be tested. Every variable will be combined with every other variable: **Predator height:** (Short (e.g., about 5ft) medium (about 5’6”), tall (about 6’)), **Proximity to bird nest:** (near (e.g., about 10 ft), moderate (about 20 ft), far (about 30 ft)), **Frequency of scanning behavior** (seldom (e.g., about once per 30 seconds), moderate (about three times per 30 seconds), frequent (about six times per 30 seconds)). (This description would consist of 3 x 3 x 3 conditions, or 27 total conditions. Other model answers may include fewer variables and levels, for example, only 2 levels of nest proximity and 2 levels of scanning frequency, for a total of 4 conditions). The bird’s calling response (or departure from nest, or feigning a broken wing response, or squeak response, e.g.) will be measured by videotaping bird and timing the latency to begin a calling response from the time the predator begins walking near the nest. Measurements will be taken for (e.g.) a two-minute time period. Bird’s responses will be potentially reduced to a categorical response if bird responses are sparse and reduced precision results in no loss of information. **Supported outcome:** Bird’s response increased in frequency (or intensity, etc.) as scanning frequency and proximity increased. **Potential confounds:** the physical appearance of the human predator may be a confound if not held constant across predator height levels or differences in predator walking style (i.e., some predators walk more heavily than others), predator gender, etc.
4. Acknowledge the possibility of alternative accounts of events and properties and judge their relative plausibility.

<table>
<thead>
<tr>
<th>Scoring Guide</th>
<th>(0-1) Not Competent</th>
<th>(1-2) Minimally Competent</th>
<th>(3-4) Competent</th>
<th>(5-6) Highly Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Acknowledge the possibility of alternative accounts of events and properties and judge their relative plausibility.</td>
<td>Fails to address how or why the choice indicates an alternative explanation of the behavior.</td>
<td></td>
<td></td>
<td>Describes how or why the choice indicates an alternative explanation of the behavior.</td>
</tr>
</tbody>
</table>

**Clarity of writing:** Competent responses are generally clearly written using correct English. Minimally competent or not competent writing is that which fails to address the question posed, or is unclear, unorganized or contains errors of English usage, mechanics, or grammar that impede the evaluator’s comprehension of the response.

Competent ☐ Minimally competent ☐ Not Competent ☐

**One example of a highly competent response:** Answer B does not counter the suggestion that bait fishing demonstrates reasoning, because whether or not it is actually unusual, the behavior shows what may be considered reasoning. For C, if the discovery is accidental but then repeated due to its success, this still arguably supports an ability to reason. For D, communication is not really relevant to this situation. Answer A is the best statement which counters the suggestion that reasoning is demonstrated because if the birds, in general, are able to reason bait fishing would presumably spread in usage, at least among birds in the same area.

A highly competent response need not come to any these exact conclusions, but should make a point and support it.
Question #5  SCORE (0-6) ______________

<table>
<thead>
<tr>
<th>Scoring Guide</th>
<th>(0-1) Not Competent</th>
<th>(1-2) Minimally Competent</th>
<th>(3-4) Competent</th>
<th>(5-6) Highly Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Identify sources of error in scientific investigation, including errors of measurement and ambiguity of judgment, or unsound logic.</td>
<td>No rationale for choice.</td>
<td></td>
<td></td>
<td>Addresses question; Reply makes clear how the proposed confounding factor can cause variation in MS rates independent of latitude or can complicate measurements.</td>
</tr>
</tbody>
</table>

**Clarity of writing:** Competent responses are generally clearly written using correct English. Minimally competent or not competent writing is that which fails to address the question posed, or is unclear, unorganized or contains errors of English usage, mechanics, or grammar that impede the evaluator’s comprehension of the response.

Competent □ Minimally competent □ Not Competent □

**Some examples of confounding factors:**
- Diet changes with latitude because of differences in locally available foods or other changes in culture. If populations in high latitudes consume foods rich in vitamin D, this would weaken the observed latitude – D deficiency correlation.
- Samples from the populations must be sex-specific or else introduce sex as a variable; if the high latitude population sample is weighted by the less susceptible sex, this would weaken the observed correlation.
- Exposure to sun during the long summer at northern latitudes may be protective through the dark winter; the total annual exposure may not be so important as some minimum dose.
- If the population is mobile a reliable estimate of exposure to sunlight cannot be inferred by subject location when interviewed.
- Current latitude may not bear on the degree of exposure at a critical time in the course of the illness.
Question #6  SCORE (0-6) _______________

<table>
<thead>
<tr>
<th>Scoring Guide</th>
<th>(0-1) Not Competent</th>
<th>(1-2) Minimally Competent</th>
<th>(3-4) Competent</th>
<th>(5-6) Highly Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Infers valid information from quantitative data presented in graphical form</td>
<td>Does not interpret graph correctly. Does not provide a recommendation and whether it would change.</td>
<td></td>
<td></td>
<td>Describes first recommendation and whether it would change. Reply demonstrates clear and correct interpretation of the graph.</td>
</tr>
</tbody>
</table>

**Clarity of writing:** Competent responses are generally clearly written using correct English. Minimally competent or not competent writing is that which fails to address the question posed, or is unclear, unorganized or contains errors of English usage, mechanics, or grammar that impede the evaluator’s comprehension of the response.

Competent ☐ Minimally competent ☐ Not Competent ☐

**One example of a highly competent response.** According to the graph, the athlete should take 20,000 IU of Vitamin D3 for at least 6 weeks to reach 75 nmol/l. At this length of treatment, serum level for participants at that dose is significantly greater than baseline levels, and can reach the optimal level determined by the coach. 40,000 IU D3 was also significant at 6-weeks, but reached serum levels which are higher than the coach believes is ideal. The athlete should avoid the placebo, which was significantly lower. If the optimal level, on the other hand, is 100 nmol/l, then the athlete should take 40,000 IU for six weeks.
Question #7  SCORE (0-6) ________________

<table>
<thead>
<tr>
<th>Scoring Guide</th>
<th>(0-1) Not Competent</th>
<th>(1-2) Minimally Competent</th>
<th>(3-4) Competent</th>
<th>(5-6) Highly Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Acknowledge the possibility of alternative accounts of events and properties and judge their relative plausibility.</td>
<td>Does not explain why the two chosen are problematic for public health.</td>
<td></td>
<td></td>
<td>Explains why the two chosen are problematic for public health.</td>
</tr>
</tbody>
</table>

**Clarity of writing:** Competent responses are generally clearly written using correct English. Minimally competent or not competent writing is that which fails to address the question posed, or is unclear, unorganized or contains errors of English usage, mechanics, or grammar that impede the evaluator’s comprehension of the response.

Competent □ Minimally competent □ Not Competent □

**Possible Answers:** (A) seems most urgent since toxicity and the potential for major health consequences must be considered. The public health campaign should emphasize that “more = better” does not apply to vitamin D. (B) is part of the usual uncertainty typical of dosage judgments; its significance depends on the seriousness of any discrepancy. (C) is important to the degree interaction is a large effect, and is a matter for evaluation and if need be container labeling. (D) (like A-C) requires some sort of understanding of the range of safe/beneficial doses. (E) becomes an issue only if customers are encouraged to take huge doses.

A highly competent response need not come to any these exact conclusions, but should make a point and support it.
Question #8  SCORE (0-12) _______________

<table>
<thead>
<tr>
<th>Scoring Guide</th>
<th>(0-2) Not Competent</th>
<th>(3-5) Minimally Competent</th>
<th>(6-9) Competent</th>
<th>(10-12) Highly Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Display a grasp of experimental design, including the notion of the control and the idea of statistical significance.</td>
<td>-Does not discuss measurement of cancer rates in two groups. -No mention of need for representative sample. -Does not consider possible confounding factors. -Does not mention possible relationship link between some agent and cancer rate.</td>
<td></td>
<td></td>
<td>-Notes need for data analysis comparing cancer rates. -Discusses representative sample. -Cites three variables which need to be measured and three confounding factors. -Notes possible alternative causes of cancer.</td>
</tr>
</tbody>
</table>

Clarity of writing: Competent responses are generally clearly written using correct English. Minimally competent or not competent writing is that which fails to address the question posed, or is unclear, unorganized or contains errors of English usage, mechanics, or grammar that impede the evaluator’s comprehension of the response.

Competent □  Minimally competent □  Not Competent □

One example of a highly competent response: Two groups of individuals need to be available for the subject pool — coffee drinkers and non-coffee drinkers. The sample should be representative of all adults and of sufficient size to detect any effects that might be present but not so large that minor effects reach the preset criterion for statistical significance. Subjects should be randomly selected from each group. A health assessment including family history should be obtained from each Subject. For the coffee drinkers, daily logs of the amount of coffee and type of coffee (regular vs. decaf) consumed must be maintained. For all Subjects, records of food consumed, smoking behavior, alcohol consumption, and other potential carcinogenic factors should be obtained. Annual (at least) medical examinations need be obtained for all Subjects and the incidents of pancreatic cancer recorded. The study should be of sufficient duration to allow for the development of cancer. Factors to consider should also include age, race, and gender of Subjects. Subjects in the non-coffee drinking group must immediately report the consumption of coffee.
Appendix 6: Mapping of 2013 Revised Test Questions to Learning Outcomes

Scientific Reasoning 2013 Test

One who can be considered competent in scientific reasoning:

1. Understands that while scientific statements are in principle tentative, criteria exist by which their validity can be judged, including consistency with the body of scientific theory, method, and knowledge; (“Criteria”)
2. Displays a grasp of experimental design, including the notion of the control and the idea of statistical significance of differences in outcome; (“Design”)
3. Infers valid information from quantitative data presented in graphical form; (“Graphic Information”)
4. Acknowledges the possibility of alternative accounts of events and properties and judges their relative plausibility; (“Alternatives”)
5. Identifies sources of error in scientific investigation, including errors of measurement and ambiguity of judgment, or unsound logic; (“Errors”)

<table>
<thead>
<tr>
<th>Learning Outcome→ Test Question ↓</th>
<th>Criteria</th>
<th>Design</th>
<th>Graphic Information</th>
<th>Alternatives</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – Kohler’s chimps</td>
<td>X</td>
<td></td>
<td></td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>2 – Tolman’s rats</td>
<td>X</td>
<td></td>
<td></td>
<td>O</td>
<td></td>
</tr>
<tr>
<td>3 – Design bird experiment</td>
<td></td>
<td>X</td>
<td></td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>4 – Bird bait fishing</td>
<td>O</td>
<td></td>
<td></td>
<td>X</td>
<td>O</td>
</tr>
<tr>
<td>5 – Vit D, MS, and Latitude</td>
<td>O</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 – Vit D serum level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>7 – Vit D public health issue</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8 – Design coffee experiment</td>
<td>O</td>
<td>X</td>
<td></td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>Distribution</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Evaluator rated each question as to how strongly it assessed each learning outcome. For all questions, except one, one learning outcome was the clear “winner” and this is marked with an X. Learning outcomes marked with a O were still scored as assessing the learning outcome, but less so, than the “winner”.
Appendix 7: Representativeness of Sample

Prior to analysis of the results, the undergraduate samples were corrected for overrepresentation by some disciplines. Fourteen (14) students from overrepresented disciplines (Science, Social Science, Engineering, Education, and Nursing) were randomly removed for a final fourth-year sample size of 156 (Table 1). Sixty-one (61) students from overrepresented disciplines (all but Humanities/Fine Arts and Architecture) were randomly removed for a final first-year sample size of 109 (Table 2).

Table 1: Determination of Fourth-Year Sample by Discipline

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Count</th>
<th>Percent</th>
<th>4th-year population percentage</th>
<th>Final 4th-year sample n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>4</td>
<td>2.4%</td>
<td>2.4%</td>
<td>4</td>
</tr>
<tr>
<td>CLAS Hum/FA</td>
<td>34</td>
<td>20.0%</td>
<td>21.5%</td>
<td>34</td>
</tr>
<tr>
<td>CLAS Social Science</td>
<td>39</td>
<td>22.9%</td>
<td>21.5%</td>
<td>35</td>
</tr>
<tr>
<td>CLAS Science</td>
<td>35</td>
<td>20.6%</td>
<td>22.1%</td>
<td>34</td>
</tr>
<tr>
<td>Commerce</td>
<td>14</td>
<td>8.2%</td>
<td>9.2%</td>
<td>14</td>
</tr>
<tr>
<td>Education</td>
<td>5</td>
<td>2.9%</td>
<td>1.8%</td>
<td>3</td>
</tr>
<tr>
<td>Engineering</td>
<td>30</td>
<td>17.6%</td>
<td>15.8%</td>
<td>25</td>
</tr>
<tr>
<td>Nursing</td>
<td>6</td>
<td>3.5%</td>
<td>2.8%</td>
<td>4</td>
</tr>
<tr>
<td>SCPS</td>
<td>3</td>
<td>1.8%</td>
<td>3.0%</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>170</td>
<td>100.0%</td>
<td>100.0%</td>
<td><strong>156</strong></td>
</tr>
</tbody>
</table>

Table 2: Determination of First-Year Sample by Discipline

<table>
<thead>
<tr>
<th>Discipline</th>
<th>Count</th>
<th>Percent</th>
<th>4th-year population percentage</th>
<th>Final 1st-year sample n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architecture</td>
<td>3</td>
<td>1.8%</td>
<td>2.4%</td>
<td>3</td>
</tr>
<tr>
<td>CLAS Hum/FA*</td>
<td>24</td>
<td>14.1%</td>
<td>22.2%</td>
<td>24</td>
</tr>
<tr>
<td>CLAS Social Science*</td>
<td>32</td>
<td>18.8%</td>
<td>22.1%</td>
<td>24</td>
</tr>
<tr>
<td>CLAS Science*</td>
<td>57</td>
<td>33.5%</td>
<td>22.8%</td>
<td>25</td>
</tr>
<tr>
<td>Commerce*</td>
<td>18</td>
<td>10.6%</td>
<td>9.5%</td>
<td>10</td>
</tr>
<tr>
<td>Education</td>
<td>6</td>
<td>3.5%</td>
<td>1.8%</td>
<td>2</td>
</tr>
<tr>
<td>Engineering</td>
<td>22</td>
<td>12.9%</td>
<td>16.3%</td>
<td>18</td>
</tr>
<tr>
<td>Nursing</td>
<td>5</td>
<td>2.9%</td>
<td>2.9%</td>
<td>3</td>
</tr>
<tr>
<td>Unknown</td>
<td>3</td>
<td>1.8%</td>
<td>0.0%</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>170</td>
<td>100.0%</td>
<td>100.0%</td>
<td><strong>109</strong></td>
</tr>
</tbody>
</table>
The sample of first- and fourth-years was compared to the population to determine any differences in measures of academic potential and success: current GPA, and SAT Verbal, Math, and Writing scores. Among first-years, there were no statistically significant differences between the sample and the first-year population on any academic indicators. But among fourth-years, the sample showed statistically significantly higher GPA and SAT scores, although the substantive difference was slight (effect sizes were .03 or less) (Table 3).

Table 3: Indicators of Representativeness of First- and Fourth-Year Samples

<table>
<thead>
<tr>
<th></th>
<th>1st-year Sample Mean</th>
<th>1st year Population Mean</th>
<th>4th-year Sample Mean</th>
<th>4th year Population Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>3.39</td>
<td>3.32</td>
<td>3.37</td>
<td>3.28</td>
</tr>
<tr>
<td>SAT Verbal</td>
<td>676</td>
<td>665</td>
<td>664</td>
<td>648</td>
</tr>
<tr>
<td>SAT Math</td>
<td>684</td>
<td>684</td>
<td>676</td>
<td>665</td>
</tr>
<tr>
<td>SAT Writing</td>
<td>681</td>
<td>674</td>
<td>659</td>
<td>653</td>
</tr>
<tr>
<td>N</td>
<td>109</td>
<td>3053</td>
<td>156</td>
<td>3515</td>
</tr>
</tbody>
</table>

Female students were over-represented in the sample. While the student population was 55 percent female and 45 percent male, the sample was 67 percent female and 33 percent male. This overrepresentation of females is not unique to these competency assessments as female students are more likely to respond to requests for participation in assessments or surveys. This gender differential is not of serious concern for this assessment, however, as gender differences were not revealed in performance on any test question.
Appendix 8: Scoring Reliability

Fifteen faculty and graduate students scored the scientific reasoning tests over a four-day workshop in May 2013. Raters were unaware of which tests had been completed by first-years, fourth-years, or graduate students. Papers were scored according to rubrics, and norming sessions were conducted prior to scoring. Each short answer question was scored on a scale of one to six, and the experimental design questions were scored one to twelve (See Appendix 5 for rubrics). In addition, raters were asked to score answers according to how well the student organized and expressed his/her response (writing score). Each student paper was scored by two raters, with the final score being the average of the two ratings.

In past competency assessments, where the scoring is done on a one to four scale, scoring reliability is considered acceptable if no more than 10 percent of ratings by evaluators differ by more than one point. In the assessment of short answer questions, while the scoring range is larger (one to six), the same reliability standard was applied. For short-answer questions, 14 percent differed by more than one point on a six-point scale (11 percent by two points and 3 percent by more than two points). The majority of ratings were either equal or differed by just one point (Table 1).

### Table 1: Scoring reliability for short-answer questions

<table>
<thead>
<tr>
<th>Q1 - Criteria to judge scientific statements</th>
<th>Q2 - Criteria to judge scientific statements</th>
<th>Q4 - Alternative Accounts</th>
<th>Q5 - Sources of Error</th>
<th>Q6 - Graph</th>
<th>Q7 - Alternative Accounts</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal Rating</td>
<td>44%</td>
<td>48%</td>
<td>37%</td>
<td>42%</td>
<td>46%</td>
<td>40%</td>
</tr>
<tr>
<td>One point difference</td>
<td>41%</td>
<td>37%</td>
<td>47%</td>
<td>44%</td>
<td>45%</td>
<td>45%</td>
</tr>
<tr>
<td>Two points difference</td>
<td>12%</td>
<td>11%</td>
<td>12%</td>
<td>12%</td>
<td>7%</td>
<td>12%</td>
</tr>
<tr>
<td>More than two points difference</td>
<td>4%</td>
<td>4%</td>
<td>4%</td>
<td>2%</td>
<td>2%</td>
<td>3%</td>
</tr>
</tbody>
</table>

The scoring range for the experimental design questions was twice that of the short-answer questions (1 to 12). Using the same standard applied in short-answer assessments, but simply doubling it results in acceptable reliability if not more than ten percent of ratings differ by more than two points. The scoring of the experimental design questions met this standard. Ten percent of ratings differed by more than two points, 27% were equal, and 63% differed by one or two points (Table 2).
Table 2: Scoring reliability for experimental design questions

<table>
<thead>
<tr>
<th>Category</th>
<th>Q3 - Experimental Design</th>
<th>Q8 - Experimental Design</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equal Rating</td>
<td>28%</td>
<td>26%</td>
<td>27%</td>
</tr>
<tr>
<td>One or two points difference</td>
<td>60%</td>
<td>65%</td>
<td>63%</td>
</tr>
<tr>
<td>Three or four points difference</td>
<td>10%</td>
<td>7%</td>
<td>8%</td>
</tr>
<tr>
<td>More than four points difference</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
</tr>
</tbody>
</table>
Appendix 9: Frequency Distributions for Each Question on Revised Test

**Question 1 - Criteria to Judge Scientific Statements**

Difference in means not statistically significant

**Question 2 - Criteria to Judge Scientific Statements**

Difference in means not statistically significant
Differences in means significant at p<.001

Differences in means significant at p<.01
Differences in means significant at $p<.05$

Differences in means significant at $p<.05$
Question 7 - Alternative Accounts

Differences in means not statistically significant

Question 8 - Observational Design

Differences in means significant at p<.01
Appendix 10: Comparison of Three Questions: 2010 vs. 2013

Fulfilling one of the 2009-2010 recommendations, the 2012-2013 Committee revised the test, primarily by deleting questions to reduce time required to complete the test. In addition, however, changes were also made to the test instructions and scoring rubrics. While a few questions remained largely intact, time to complete the answer was reduced, format for entering answers was changed, and the scoring was altered.

Three questions remained intact—Q1, Q2, and Q3—although test instructions and time to complete answers were changed.

Table 1: Comparison of Three Questions, 2010 vs 2013

<table>
<thead>
<tr>
<th>Q-Type of Question/Topic</th>
<th>Task</th>
<th>Suggested time</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1- Short Essay/Kohler’s chimps</td>
<td><em>In a few short sentences, explain your choice among the alternatives.</em></td>
<td>2010- no instruction</td>
<td>2010- 3 outcomes plus Writing: (1=Not Competent, 2=Minimally Competent, 3=Competent, 4=Highly Competent)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2013- no more than 5 minutes</td>
<td>2013- 1 outcome (0-1 Not Competent; (1-2)-Min Competent; (3-4) Competent, (5-6) Highly Competent)</td>
</tr>
<tr>
<td>Q2- Short Essay/Kohler vs. Tolman</td>
<td><em>In a few short sentences, explain your choice</em></td>
<td>2010- no instruction</td>
<td>2010- 4 outcomes plus Writing: (1=Not Competent, 2=Minimally Competent, 3=Competent, 4=Highly Competent)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2013- no more than 5 minutes</td>
<td>2013- 1 outcome (0-1 Not Competent; (1-2)-Min Competent; (3-4) Competent, (5-6) Highly Competent)</td>
</tr>
<tr>
<td>Q3-Experimental Design Essay/Bird-Predator Conditions</td>
<td><em>Devise a controlled experiment...in 1-2 paragraphs, describe the experiment</em></td>
<td>2010: 10-15 minutes</td>
<td>2010- 3 outcomes plus Writing: (1=Not Competent, 2=Minimally Competent, 3=Competent, 4=Highly Competent)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2013: 10 minutes</td>
<td>2013-1 Outcome (0-2) Not Competent; (3-5) Min Competent, (6-9) Competent, (10-12) Highly Competent</td>
</tr>
</tbody>
</table>
As such, they offered a limited opportunity to compare 2013 results with those from 2009. Figure 1 and Table 2 show percent of possible scores attained for undergraduates in 2009 compared to undergraduates in 2013. Differences are statistically significant, with effect sizes in the small to medium range.

Figure 1: Mean Percent of Possible Points Earned on Q1, Q2, and Q3: Original vs. Revised Test

Table 2: Percent of Possible Points Earned on Q1, Q2, and Q3: Original vs. Revised Test

<table>
<thead>
<tr>
<th></th>
<th>Original Test: 2009 1st and 4th years</th>
<th>Revised Test: 2013 1st and 4th years</th>
<th>sig</th>
<th>effect size of the difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 Judge Scientific Statements</td>
<td>57%</td>
<td>36%</td>
<td>p&lt;.001</td>
<td>0.18</td>
</tr>
<tr>
<td>Q2 Judge Scientific Statements</td>
<td>60%</td>
<td>54%</td>
<td>p&lt;.01</td>
<td>0.02</td>
</tr>
<tr>
<td>Q3 Experimental Design</td>
<td>62%</td>
<td>41%</td>
<td>p&lt;.001</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Comparing only fourth-years in 2012-13 (those who took the original test as part of the longitudinal assessment vs. those who took the revised test as part of the competency assessment), the differences again are statistically significant but with very small effect sizes (Figure 2, Table 3). In both cases, however, the students answering the three questions on the revised test scored significantly (and in some cases substantially) lower than students who answered the three questions on the original test.
Question 3, an experimental design question, provides the clearest picture. In 2009, fourth-year students’ scores ranged from about 40 percent to nearly 100 percent, with an average of 62% (Figure 3). In 2013, however, the distribution shifted sharply to the lower end with scores ranging from less than 20% to 83%, averaging 41% (Figure 4).
Figure 3: For Q3: 2013 Performance by Fourth-years given 10-15 Minutes and Scored on X Outcomes (Original Test)

Figure 4: For Q3: 2013 Performance by Fourth-years given 10 Minutes and Scored on One Outcome (Revised Test)

A likely contributor to this difference was the change in time allotted for the question, which was reduced from 10-15 minutes (original test) to 10 minutes (revised test). It is reasonable to conclude that student performance may have been adversely affected—more students scored lower and fewer scored higher. With less time to think and write, students were less likely to excel. In addition, the scoring for this question changed. In 2010 three outcomes (plus writing) were assessed on a scale from 1 to 4
whereas in 2013, only one outcome was assessed, but on a scale from 1 to 6. It may be that the raters graded harder on the sole outcome in 2013 than they did on each of the multiple outcomes in 2010.

In its current form, this question may well test students’ ability to think and write quickly as much as their ability to design an experiment. Quick thinking, however, was not one of the learning outcomes. As an unintended consequence of the test revision, it may have been inserted as a confounding variable in the assessment.

As such, the reduction in time allotted to answer the question and the change in scoring may have made the test less able to solicit and detect the full range of competency among undergraduates, a Type II error. Perhaps because graduate students, especially those in the sciences, are more adept at scientific reasoning, they could score higher regardless of the time constraint and scoring changes.