A LIVELY ELECTRONIC COMPENDIUM OF RESEARCH, NEWS, RESOURCES, AND OPINION

Astronomy Education Review

2013, AER, 12(1), 010113, http://dx.doi.org/10.3847/AER2012027

Multiple Choice Testing for Introductory Astronomy: Design Theory Using Bloom's Taxonomy

Arthur Young
San Diego State University, San Diego, California 92182
Stephen J. Shawl
University of Kansas, Lawrence, Kansas 66045
Received: 07/6/13, Accepted: 07/29/13, Published: 09/13/13

Abstract

Professors who teach introductory astronomy to students not majoring in science desire them to comprehend the concepts and theories that form the basis of the science. They are usually less concerned about the myriad of detailed facts and information that accompanies the science. As such, professors prefer to test the students for such comprehension. The multiple choice format for examinations is often excluded since it appears to focus only upon factual information. This paper proposes to show that the multiple choice format can be used to create non-trivial examinations that test for higher-order thinking. The paper shows, with numerous examples, how to design such questions within the didactic framework of Bloom's taxonomy. Following a discussion of how the taxonomy relates to goals and objectives in teaching astronomy, the paper proceeds to focus upon each of the taxonomic categories with examples of sophisticated questions about specific astronomical concepts. The goal is to explicate the design theory so that instructors can create unlimited numbers of questions for their own courses. As such, it is not a research paper but, rather, one to provide working instructors with helpful hints for improving their testing. Included at the end of the paper is an algorithm for the construction of a grade curve, and some discussion of using the statistical analysis of the examination questions to evaluate the performance of individual students and to improve the test questions themselves.

1. INTRODUCTION

1.1. The Tyranny of Testing

Following a long series of articles in *The American Scholar, Harper's Magazine, and Physics Today,* the eminent physicist and scholar Banesh Hoffmann published a book titled, *"The Tyranny of Testing"* (Hoffmann 1962). In his typically cogent and forceful manner, Hoffmann set out in that monograph the entirety of his strong objections to the widespread use of the multiple choice form of testing and evaluation of human skills, knowledge, performance, and intellect. Anyone who has ever used such a method of testing, or who is contemplating such a venture, might enjoy reading Hoffmann's sublimely introspective analysis.

Hoffmann's positions are well documented and well thought through and expressed. His thesis, in very broad overview, is that the multiple choice format does not actually test most of what its users purport to be testing, and often it puts the most able students at a disadvantage relative to their clever, but less able, peers. The scholarly and general condemnation of such testing by Hoffmann must be carefully examined, for if it is universally true, then such examinations should be banished (perhaps Baneshed) from the academic evaluation process. A somewhat more specific synopsis of Hoffmann's viewpoint is that multiple choice tests fail to test the depth of native human intelligence (whatever that really is) or of specific human aptitudes for such domains as science, mathematics, or writing. He argues that they especially fail to evaluate human creativity in any of its variegated

forms. Even more mundane matters such as critical thinking, reading comprehension, and communication skills are tested only at a superficial level, placing the finest and deepest thinkers at a relative disadvantage. In a summarized form, his assertions provide support by noting the following facts about the multiple choice format.

1. The format denies creative thinkers any opportunity to demonstrate their originality.
2. Shrewd but superficial thinkers have a distinct advantage over deeper thinkers who often discover in questions subtle ambiguities that were not noticed by the examiner.
3. The format denies the opportunity to elucidate the quality of thought that led to a correct answer (i.e., convoluted logic; penetrating analysis; or shrewd guessing).
4. The best-prepared students often see deeper into questions than the examiner intended and thus become confused when none of the choices seem to be correct, or more than one could be correct.
5. The format prevents the examiner from asking the most deep and searching questions that are possible.
6. Awkward grammars or a poor choice of a word or phrase can introduce unintended ambiguity into an otherwise reasonable question.
7. The format encourages and may even reward guessing.
8. Deep understanding of concepts may not be tested and may even be lacking in those persons who score very well.
9. In cases where more than one answer could be equally correct, the student must attempt to probe the mind and the intent of the examiner rather than the subject.

From such an array of "faults," one might conclude that it is difficult to see any redeeming qualities to multiple choice examinations. As we hope will be clear, such a conclusion would be wrong.

Almost exclusively throughout his monograph, Hoffmann's criticisms are directed at the various products of the Educational Testing Service (E.T.S.) and the College Entrance Examination Board (C.E.E.B.) with special attention to the Scholastic Aptitude Test (S.A.T.) and the Graduate Record Examination (G.R.E.). Also among his targets are various I.Q. tests, the National Merit Scholarship Corporation, and various military and corporate testing for job aptitudes and promotions. Conspicuous by its absence is any direct critique of the use of multiple choice examinations for the evaluation of performance in specific courses, but there is little doubt of what Hoffmann's view of that would be.

A careful analysis of Hoffmann's critiques reveals that his analysis is both astute and largely correct. However, the fallacy of the testing methodology comes not from some inherent flaw in the multiple choice format, but rather in its indiscriminate use to test attributes for which it is clearly unsuited. Neither Hoffmann nor anyone else has offered a convincingly effective method for assessing native intelligence, scholastic (or other) aptitudes, or general intellectual achievement. A case might even be made that there is no effective way to quantify and measure such intangibles. Hoffmann's analysis leaves little doubt that the multiple choice examination format falls far short of achieving such goals, regardless of whether they are ultimately achievable or not.

In the discussion that follows, we maintain that the multiple choice format can be used effectively for the limited purpose of evaluation and certification of modest but measurable amounts of intellectual growth in sharply defined domains of knowledge. A review of guidelines for writing multiple choice questions is given by Haladyna, Downing, and Rodriguez (2002).

Few higher education faculty members are trained in writing validated and reliable assessments. A test question has validity when shown to measure what was intended to be measured. A question has reliability if it consistently gives repeatable results under consistent conditions. Expressed in terms that a physical scientist is most familiar with, we can say that *validity* corresponds to *accuracy* while *reliability* corresponds to *precision*. Just as one can have an astronomical measurement that is precise but not accurate, one can have a test item that is valid but not reliable.

Significant efforts have been made in recent year to produce reliable and valid concept inventories and surveys. Given their inclusion of questions that require them to be both valid and reliable, these papers provide an understanding of the process involved in obtaining validity and reliability. Examples are the Astronomy Diagnostic Test (Hufnagel 2001), light and spectroscopy concept inventory (Bardar *et al*. 2006), cosmology surveys (Wallace *et al*. 2011), properties of stars concept inventory (Bailey 2007), and a review of research in Ph.D. dissertations (Slater 2008).

1.2. General Education Science

In spite of differences in rhetoric among institutions, the basic goals of science courses in the liberal arts curriculum differ very little. Teaching methods, course contents, and styles differ more than do the basic objectives for such courses. In particular, there is no intention to teach science as if to a future professional, and thus there is no expectation that the amount of knowledge gained or the depth of its comprehension be commensurate with that of first year science students. Thus, while exceptions exist, there is no over all need in general education courses for tests to evaluate depth of insights, problem solving competency, analytical skills, and originality since those are not always among the goals of such courses.

The best courses strive to heighten the awareness of non-science students to the world of science and its values and methodology, while giving a greater appreciation for what science is about and what it has discovered about reality. High on the list of goals in all such courses is to impart some sense of critical thinking about scientific matters (and all else) and to elucidate the facts that lay at the foundations of contemporary knowledge and understanding of the universe. Some understanding of the great principles and laws that seem to define the structure of nature is central to all such courses. (See Partridge and Greenstein 2003 for a detailed discussion of ASTRO 101 goals.)

Testing and evaluation are therefore directed at the assessment of growth in these broader domains, within the confines of the subject matter that comprises the course. The objective, multiple choice format is capable of providing such an assessment, but only if great care is taken in the creation of the questions. In particular, the overwhelming tendency for such questions to do no more than elicit useless memorized information must be avoided. It is possible, albeit challenging, to design multiple choice questions that probe for comprehension as well as knowledge, and for the limited ability to analyze and to apply knowledge that is expected of students who complete such courses.

The prevalent guide for the hierarchy of learning in contemporary education is Bloom's taxonomy (Bloom *et al*., 1956), which has six levels in order of increasing sophistication. An update to the original hierarchy was proposed by Bloom's former student (Anderson and Krathwohl 2001; Krathwohl 2002). The revised terms, in which nouns are changed to verbs, are consistent with the active learning approaches of today. The new terms are given in parentheses after Bloom's original terms, which are still very much in use today. Also, the original numbers 5 and 6 are now reversed with "create" being the highest level. Nonetheless, Bloom's original terms will be used for the rest of this paper.

1. KNOWLEDGE (remember): Information of fundamental importance, and the ability to recall it for appropriate situations. In astronomy, this translates into knowing significant observational facts (e.g., the existence and properties of the cosmic background radiation), and fundamental laws such as those of Kepler.
2. COMPREHENSION (understand): Understanding the meaning of knowledge and how it is obtained and justified. Following the previous example, this would imply an understanding of the interpretation of the cosmic background radiation and its implication for the structure of the universe; and an understanding of the implications of Kepler's laws upon the motions of orbiting objects and how those laws derive from the more general law of universal gravitation.
3. APPLICATION (apply): Ability to use knowledge in new or unfamiliar situations. Recognition, for example, of how Kepler's laws can permit the determination of the mass of a planet would constitute a demonstration of this ability by students in an introductory astronomy course.
4. ANALYSIS (analyze): An advanced form of application consisting of the ability to use knowledge to understand or to infer a new situation. Using Kepler's laws in a clever way to predict distance of an Earth satellite, whose period is that of the rotation of the Earth is an example.
5. SYNTHESIS (create): Ability to combine two or more concepts or facts to infer yet another, or to achieve analysis. An example might be making assumptions about the structure of the Milky Way galaxy combined with Kepler's laws leading to an estimate of the number of stars in the Milky Way galaxy.
6. EVALUATION (Evaluate): Judgments based on critical thinking using acquired knowledge. The mature student evaluates the contemporary obsessions with astrology, the occult, the supernatural, and Creationism. A more subtle example is the ability to recognize the difference between a strongly suggestive argument, and one that is truly compelling. Evaluation is the most sublime form of critical thinking.

Distinctions between these categories are not always sharp, nor are they independent and mutually exclusive. The ability to apply knowledge presupposes ownership of the knowledge and reasonable comprehension of it as

well. Analysis is a form of application, and synthesis usually involves some analysis. Good test questions will often combine more than one of the skills described by Bloom's taxonomy, while focusing upon one in particular. In the discussion of the examples, reference will be made wherever appropriate to particular categories in the original taxonomy.

1.3. Critical Thinking and Evaluation

Science is widely thought to epitomize rational and logical thinking, to the exclusion of the converse of those attributes. Such a parochial view of science overlooks the imaginative originality responsible for whatever could be called *great* about science. Nevertheless, students are often urged to learn some science in order to learn to think rationally and logically. One would hope that rational thinking is practiced universally in the academic community, so that a science course is not the sole exposure to such a virtue. Logic, a particular form of rational thinking, is better learned from logicians and mathematicians than from scientists who use it when it serves their purposes (and who may abandon it when it does not). However, one particular attribute of rationality that scientists develop to a high degree is *critical thinking*. The habit and the skill of critical thinking are a direct outgrowth of the intellectual competition to find the *correct* interpretation for the array of clues that nature displays about its construction. Since more than one interpretation is almost invariably possible for any set of observed phenomena, the quest for validity will always involve a critical analysis of different ideas that are put forward by original thinkers.

The training of a scientist acquaints him or her with the history of such critical thinking, and the practice of science exposes him or her to its use. The proposals and the findings of a working scientist are subjected to the critical review of his peers, and the working scientist finds himself to be a critical reviewer of his colleagues. Experienced scientists become masters of critical thinking almost inadvertently. Even though our ASTRO 101 courses are not designed to train professional scientists, nevertheless critical thinking, and in particular evaluation, should be central components of what we teach and what we test for in our introductory science courses (Partridge and Greenstein, 2003). Below are some of the attributes that a skilled *critical thinker* must possess.

1. The ability to recognize the presence of unsubstantiated assumptions in an argument; especially when the assumptions are tacit or implicit and not stated explicitly.
2. The ability to recognize ambiguities or alternatives in the interpretation of otherwise factual information.
3. Awareness of the potential for unknown or unavailable information to alter an impression or an interpretation (i.e., the tentativeness of all conclusions based entirely upon facts).
4. Recognition of fallacies in statistical reasoning (inferring) using valid statistical information.
5. Alertness to subtle selection effects in the acquisition of sampled data, and their effects upon conclusions.
6. Recognition of non-sequiturs and other forms of convoluted logic (e.g., *post hoc ergo propter hoc*) in arguments.
7. Ability to distinguish between arguments that are weak, arguments that are strong (weaker alternatives are possible), and arguments that are compelling (all reasonable alternatives are eliminated).

Testing for critical thinking skills in a general education course is difficult. The basic strategy is to present situations in which most or all of the choices are reasonable, but only one can stand the test of sharp critical thinking. These are often the most difficult questions on any test. Frequently, the history of science offers a rich field for generating such questions.

1.3.1. Examples Requiring Critical Thinking

1. In his determination of the circumference of the Earth, Eratosthenes made each of the following assumptions except one. Which one was NOT a necessary assumption?
   A. The shape of the Earth is a sphere.
   B. The distance from Syene to Alexandria is known.
   C. The distance to the Sun is much larger than the radius of the Earth.
   D. The Earth rotates on an axis.
   E. Light travels on straight lines.

2. All of the following are actually observed phenomena. Which one permits a *compelling* argument that the Earth is shaped like a sphere?
   A. Ships vanish over a distant horizon.
   B. The sky appears to turn around us.
   C. The altitude of the North Star increases as an observer travels northward.
   D. During a lunar eclipse, the Earth's shadow appears to be circular.
   E. All lunar eclipses show the Earth's shadow to be circular.

The intended answers are: 1-D; 2-E.

Evaluation is both the most sublime form of critical thinking and the most difficult to test. By evaluation we mean the use of critical thinking to exercise judgment, and often that must be subjective. A partial list of attributes for skilled evaluation includes:

1. Ability to distinguish between poor assumptions and good assumptions when making assumptions is unavoidable.
2. Ability to distinguish weak arguments from strong arguments.
3. Ability to assess likely vs. unlikely interpretations for factual information.
4. Ability to assess the reliability of factual information, considering such matters as measurement uncertainty and selection effects.

A touchstone of critical evaluation that is easily tested in the multiple choice format is the ability to distinguish between model-dependent and model-independent information. An abbreviated example of such questioning is:

> Identify which of the following factual statements is model-dependent and which is model-independent using the following code:
>
> A = model-independent
> B = model-dependent
>
> 1. The inclination of the ecliptic to the celestial equator is 23.4419°.
> 2. The Sun takes 365.2422 days to move from the Verna 1 Equinox around the ecliptic and back to the Vernal Equinox.
> 3. The temperature at the center of the Sun is 13 478 000 K.
> 4. All of the dinosaurs perished in a relatively short time that took place 65×10^6 years ago.
> 5. In February 1987, a supernova explosion was observed in the southern sky. A newspaper report quoted an astronomer who said that the energy liberated by the explosion was about 250×10^9 times the luminous energy of the Sun.

The intended answers are: 1-A; 2-A; 3-B; 4-B; 5-B.

2. CONSTRUCTION OF MULTIPLE CHOICE QUESTIONS

Professor Hoffmann's analysis casts doubt upon the presumed *objectivity* of multiple choice examinations, but their efficiency is beyond question. For professors with enormous sized classes, or many sections of smaller ones, the grading of examinations becomes a major consumer of time. For some, that may well be just part of the responsibility of teaching the course, but many others have other courses to teach and are engaged in research programs and other academic duties. An efficient testing method that does not sacrifice rigor is highly desirable.

Those who hold the opinion that only a written essay examination retains the desired rigor would do well to read Chapter 3 of Hoffmann's book before dismissing the multiple choice format *a priori*. There is a good reason why he did not choose to call his book, *The Tyranny of Multiple Choice Testing*! It is particularly ironic, if not entirely hypocritical, when faculty members who eschew the use of multiple choice examinations for their own courses utilize S.A.T. and G.R.E. test scores as virtually infallible indicators for entrance into their own programs or even as a Ph.D qualifying examination!

Good examinations are teaching and learning experiences as well as tools for evaluation. By the particular questions that are asked, the instructor tacitly communicates which things are most important. By posing questions that require application, students discover how that is done since that is not a normal part of the studying process for non-science students. Even more significantly, questions that elicit application, analysis, and

synthesis make the students aware of just how much (or little) growth they have achieved because of their passive assimilation of new knowledge. Many students, even those who attain high grade scores, do not appreciate the magnitude nor the significance of what they have learned unless they are pressed into using it by an examination that requires much more than rote responses.

In what follows, we give examples of questions that are designed to focus upon each one of the categories in Bloom's taxonomy, though not to the exclusion of all the other categories. We discuss the strategy of each question itself, and strategies for creating questions in that category. By elucidating the strategies for doing that, we hope to make that process simpler and quicker than it is reputed to be. Following those discussions, we consider the theory of creating the foils (also known as *distractors*); i.e., the incorrect responses in each question that accompany the correct one.

Unlike a first year course in physics or mathematics, our goal is not to test for a level of competency or of potential for more advanced studies yet to come. We are testing instead for expanded awareness about the nature of science as an investigatory process; for comprehension of concepts about nature; for critical thinking; and for the foundations upon which knowledge is based. We are also testing to see if new knowledge has been acquired, and if old misconceptions have been eradicated. Using Bloom's original taxonomy as an outline, we discuss the strategies for such testing under each of the categories.

2.1. Knowledge/Remember

In its broadest sense, science can be viewed as all of the activities and processes by means of which human beings journey from *awareness* to *comprehension* concerning the phenomena of nature. Science begins with awareness of the existence of physical things, of their observable properties, and of physical occurrences or processes. Therefore, the body of *knowledge* contains an enormous amount of such empirical information, without interpretation. Comprehension comes from conceptualization, a uniquely human process. The body of *knowledge* then also contains models and theories that explain and interpret the empirical information and give meaning to those things about which we are aware.

For example, we (collectively) are aware that the Earth experiences a regular and annual variation of seasons characterized by significant variations of the ambient temperature. For many students that awareness is local, coming from their own personal experience. One form of expanded awareness comes from the disclosure that the same seasonal variations that we experience in the northern hemisphere are experienced in the southern hemisphere, but out of phase by one-half year. An explanation for the seasonal variations will be connected to the behavior of the Sun, but the hemispheric dichotomy eliminates variations of the distance as an explanation. The comprehension becomes deeper when other models are incorporated to explain why the Sun behaves in the manner that causes the seasonal variations to occur. Testing for knowledge of this subject would center upon the empirical properties such as the hemispheric seasonal distinction, and the apparent motions of the Sun at various critical latitudes like the equator, the Arctic and Antarctic circles, and the poles.

Astronomers are aware that faint microwave radiation is detected coming from all directions in the sky. We are aware of many of its properties such as its spectrum, and its nearly isotropic intensity, and its small departure from perfect isotropy. All of that is empirical knowledge, but because of its enormous significance it is appropriate that our students share that knowledge, and hence there will be test questions to insure that the existence and the principal properties of this radiation is known to them. However, it is equally important that the body of knowledge that we transmit, and for which we test, includes our sense for what is the importance of this radiation for understanding our universe, and what is our current explanation for its existence and its properties.

Knowledge, as we define it for the general education course in astronomy, has two components. First is an *empirical* component that consists of awareness of one's surroundings, of their observable properties, and of phenomena that occur. Second is a *conceptual* component that consists of knowing the contemporary explanations for such properties and phenomena, even if that does not include a deep and genuine understanding of those explanations. Under this category, we test for the knowledge of explanations, whereas we test for a deeper level of comprehension in succeeding categories. If the testing is constructed properly, the grade that is achieved represents some measure of the sophistication of learning rather than just the amount.

Information is a form of knowledge, often in a very compressed way, such as a single number. One set of examples follows:

1. The (approximate) diameter of our Milky Way galaxy.
2. The (approximate) number of stars in our Milky Way galaxy.
3. The (approximate) age of our Sun.
4. The scale of the observable universe and the (approximate) distance to the most remote objects yet detected.

Yet another set of examples of information is:

1. The number of planets in our solar system.
2. The names of the planets and their order from the Sun.
3. The seven principal spectral types of the stars
4. The three principal types of galaxies.

The first set of factual information has real significance, having emerged from scientific inquiry, and it conveys some genuine knowledge about the universe in which we live. Items 2 and 3 of the second set are useless superficialities that do not have such significance. People assign names to things arbitrarily, and they typically contain no information about the entities themselves. Likewise, classes and types may seem *scientific* but they are merely arbitrary distinctions made by scientists, and they also contain little or no physical information of any particular importance. Items 1 and 4 can be viewed superficially or with great sophistication. For example, the number of planets can be determined by dynamical considerations, but it also depends on how "planet" is defined. The principal types of galaxies are certainly determined by physical conditions and the environment but can also be viewed simply as a classification to be memorized without further understanding.

When we pose questions that ask for memorized information, we are informing the students of what we think is sufficiently important that it should be committed to memory by any person who considers himself to be well informed about our science. In that same vein, we must ask ourselves what we would want our students to remember from our course five years from now and on into the future. We should keep in mind that anything that makes a profound impression (such as the size of our galaxy, or the distance to the most remote objects) is more likely to be retained for a lifetime than something as meaningless as the scrambled order of the spectral types of the stars. By keeping these simple aphorisms in mind, most of the trivial questions that typically inhabit multiple choice examinations will vanish.

Perhaps, the most important kind of knowledge consists of those things or those phenomena that constitute *evidence* for particular understanding or viewpoints currently held by the scientific community. Since educated persons do not customarily accept things on faith, the burden is upon us to show the evidence for our own understanding. (Note: the original manuscript had "beliefs" rather than "understanding." "Belief" has a certain connotation to it that is best left out of the classroom.) Therefore, the most sophisticated kind of (memorized) knowledge is that which constitutes evidence for very important concepts and viewpoints, such as for evolution, or for the expansion of the universe. An important subset of that knowledge is empirical tests, which are critical tests that discriminate between competing theories and hypotheses.

In summary, five domains of knowledge are sufficiently important to be explicitly tested in examinations. These domains are:

1. Important phenomena or properties of nature.
2. Contemporary explanations for those phenomena or properties.
3. Significant information about natural structures.
4. Evidence upon which contemporary viewpoints are based.
5. Critical (observational) tests that distinguish between competing explanations for the same phenomena.

Each of the following sample questions illustrates the testing of one or more of those domains.

1. One month from now, a constellation of stars that we see nearly overhead at 9:00 P.M. tonight will be seen at 9:00 P.M. to be
   A. nearly overhead.
   B. lower in the western sky.
   C. lower in the eastern sky.
   D. lower in the southern sky.
   E. already set below the western horizon.

2. The patterns or constellations of stars seen in each of our seasons are different. That fact demonstrates that
   A. the Earth rotates upon an axis.
   B. the Earth revolves on an orbit around the Sun.
   C. the Earth is a sphere.
   D. the Sun moves relative to the stars.
   E. there must be parallax angles for the stars.

3. The night sky is very dark. The best currently accepted explanation for that observed fact is that
   A. the Sun is below the horizon.
   B. the universe is not infinitely large.
   C. the universe is expanding.
   D. the universe is evolving.
   E. the universe has a finite age.

4. How does the age of the oldest stars that astronomers have ever found compare to the age of the oldest rocks that geologists have ever found?
   A. The oldest stars are about twice the age of the oldest rocks.
   B. The oldest stars are 10 times older than the oldest rocks.
   C. The oldest stars and the oldest rocks are about the same age.
   D. Astronomers cannot determine the age of the oldest stars.
   E. Geologists cannot determine the age of the oldest rocks.

5. Astronomers are convinced that the Sun is evolving. What is the *observed* evidence for that conviction?
   A. Changes are actually observed on the surface of the Sun.
   B. Ancient records show the Sun was different long ago.
   C. The mass of the Sun is observed to be decreasing.
   D. The Sun radiates energy that is not being replaced.
   E. 25% of the mass of the Sun is helium.

6. The redshift of the spectra of distant galaxies constitute evidence that
   A. the universe is becoming cooler.
   B. the universe is expanding.
   C. the universe is evolving.
   D. the universe is in steady state.
   E. the galaxies are getting bigger.

7. Direct observational evidence that the scale of the universe is expanding comes from
   A. the cosmic background radiation.
   B. the darkness of the night sky.
   C. the redshifts of the spectra of galaxies.
   D. the existence of quasars.
   E. the clustering of galaxies.

8. Each of the following statements is correct, but only one constitutes *evidence* that evolution has occurred in the universe. Which statement is evidence for evolution?
   A. Very distant galaxies appear smaller than nearby galaxies.
   B. Very distant galaxies appear dimmer than nearby galaxies.
   C. Ellipsoidal galaxies contain no free dust and gas.
   D. All quasars are very far away.
   E. There are more ellipsoidal galaxies than spiral galaxies.

9. A *critical test* that demonstrates that the Earth rotates upon an axis is
   A. the change of direction of a free swinging pendulum.
   B. the moving shadow on a sundial.
   C. the parallax of the stars.
   D. eclipses of the Moon.
   E. that only one hemisphere of the Moon is ever seen from the Earth.

The intended answers are:

1-B; 2-D; 3-E; 4-A; 5-D; 6-B; 7-C; 8-D; 9-A

Question (1) tests for rudimentary knowledge about the behavior of the sky. It would have been a ridiculous question to ask in ancient Athens since everyone with eyes knew how the sky behaves, even if not why it

does so. In our modern technological age, few people look directly at nature, and the motion of the sky is a surprise to most of our students and visitors to the observatory. The foils are crafted so that the better students will either select the correct answer at once, or ponder between choices B and C. Choices A and E implicitly quantify the phenomenon, and students who have actually looked at the sky will eliminate them at once. More capable students will recognize that one month generates about 30° of sky displacement and will thus eliminate those choices. A student who selects choice D reveals a more serious academic problem.

Question (2) inverts the process, stating the observed behavior, and asking for its *unambiguous* meaning. Even some of the better students slip and select choice B, the commonly accepted *explanation* for the phenomenon that is described. The wording is careful and unambiguous; if that observed behavior could be understood *only* by orbital motion of the Earth, the geocentric astronomy of the ancients would never have been conceived. Although we have classified this as a KNOWLEDGE question, some critical thinking is needed to distinguish the subtle difference between choices B and D.

Question (3) is esoteric, and one would expect that it can be answered successfully only by persons who have taken a modern course in astronomy. A student who answers it correctly might not understand the subtle reasoning that makes choice E the only correct response. Hoffmann might criticize the question for that weakness, but it was never intended to test that deeply. It tests for esoteric *knowledge*, not for esoteric *comprehension*. There are many variations on the phrasing of the question so that it can be used with slightly different nuance on different examinations. For example, question 3 might be written:

3. The darkness of the night sky is taken by modern astronomers to be evidence that
   (The selection of responses is the same as in Question 3 above.)

3. Astronomers now understand that the universe is not infinitely old. Which of the following observed properties of the universe is evidence for that conviction?
   A. The redshift of the spectra of distant galaxies.
   B. Galaxies exist in clusters.
   C. The night sky is very dark.
   D. Most of the mass of the universe is hydrogen.
   E. Our Sun is still on the main sequence.

The intended answer is C.

Question (4) is intended to inform the students who memorized the age of the oldest known stars that such information is more meaningful when it can be put into a broader context.

Although question (5) is categorized as a KNOWLEDGE question, the selection of response D requires some comprehension of what is actually meant by evolution. A student who selects response D may not know the intricate details of the predictions of solar evolution, but demonstrates instead a mature realization of what evolution means, and what conditions insure that it must occur. It is far more likely, and certainly more valuable, that such an insight will be retained long after the complex details of solar interior evolution have faded from memory.

Questions (6) and (7) are two versions of the same question with a different twist. They are not intended for use on the same examination. There is a conceptual symmetry between knowing what evidence supports a particular viewpoint (question 7), and what viewpoint is supported by a particular piece of evidence (question 6). Students in a general education science course should become adept at both of those versions.

Question (8) is a somewhat more subtle variance of the idea that is tested by question (5). Once students recognize that evolution of any physical system is characterized by irreversible change, then evidence for evolution is going to be found either in a record of the past that can be compared with the present, or in some process that cannot reverse. Question (5) tested for recognition of an irreversible process, and question (8) tests for recognition of the difference from a former time. Although we classify question (8) as one for KNOWLEDGE, it requires some synthesis as well, which shows that the taxonomy is complex and not simple to apply. The students should *know* that quasars have unique physical properties that differ markedly from ordinary galaxies (even if they do not recall exactly what those differences are). The more astute students should recognize that “far away” translates into “long ago.” Putting these ideas together (synthesis) means that the absence of nearby quasars is a signal from the universe that it was once different from the way it is now—the signature of evolution.

Choices C, D, and E in question (9) are all *critical tests,* but for concepts other than the rotation of the Earth. The student who mindlessly memorizes all things in the course notes that were ever called "critical tests" finds himself confronted with four of them, when only one is the right one for this circumstance.

2.2. Comprehension/Understand

Those who teach elementary science courses usually wish to place a greater emphasis upon COMPREHENSION than on any of the other categories of Bloom's taxonomy. The reason is because knowledge without understanding is of little use, and the more sophisticated attributes of application and analysis generally require more mathematical training and more advanced scientific course work. The distinctions between comprehension, application, and analysis are not at all sharp, and the overlap is considerable. Nevertheless, it is possible to sort them out in the testing arena, though never exclusively. The principal strategy to test for understanding of concepts is to require the student to provide an *explanation* for why something happens the way it does. Comprehension is not demonstrated if it is possible to remember an explanation given in a textbook or a class discussion. Therefore, it is necessary to confront the student with unfamiliar situations that can be explained in terms of concepts that are supposedly comprehended.

Understanding (of concepts, models, or laws) can also be demonstrated by correctly associating a phenomenon with a physical law or with a specific model that explains it, with the caveat that it not be possible to have memorized that association. A variation that permits use of mathematics is to correctly associate a phenomenon with an equation that expresses the idea that explains the phenomenon. The latter can be done either by embedding the equation in the choices along with other, irrelevant equations, or by providing a list of equations with the examination and requesting that the appropriate one be selected. The latter technique has the additional advantage of discouraging students from memorizing equations since they know that all of them will be listed on the examination, and encouraging them to focus upon the meaning of the equations.

The following examples show all of these strategies.

1. When you switch off the lights in your room at night, the walls, the ceiling, and the floor are at a temperature of 300 K. Why are you not dazzled by the radiation that they emit?
   A. Because they do not emit any radiation.
   B. Because they are not anything like black bodies.
   C. Nothing can radiate below a temperature of 3000 K.
   D. Human eyes cannot see infrared radiation.
   E. Human eyes cannot see ultraviolet radiation.
2. The absorption lines of calcium in the spectrum of our Sun's atmosphere are much darker than those of hydrogen. The explanation for this is that
   A. the Sun is a very unusual star.
   B. there is more calcium than hydrogen in the Sun.
   C. most of the hydrogen is ionized.
   D. the hydrogen has been converted to helium.
   E. most of the hydrogen is in the ground state.
3. The shadow of a flagpole in sunlight does not appear to have a sharp edge (boundary). The reason for that is
   A. light does not travel on exactly straight lines.
   B. the Earth's atmosphere refracts the light.
   C. the Sun is not a point source of light.
   D. scattering by the Earth's atmosphere diffuses it.
   E. flagpoles are round and thus have no sharp edge.
4. Pushing a shopping cart is much easier to do than pushing an automobile. Write the letter corresponding to the equation on the back page that expresses this fact.
5. Halley's comet orbits the Sun on a very elliptical orbit with a period of 75 years. When it was near the Sun in 1985 it spent only 5 months there. That is a direct consequence of
   A. the comet being in an elliptical orbit (Kepler's I law).
   B. objects near the Sun moving faster than those far from the Sun (Kepler's II law).
   C. objects whose average distance is father from the Sun taking longer to complete its orbit (Kepler's III law).

D. the basic property of all conic sections.
E. perturbations by the inner planets.

6. Which one of the following statistical samples could be considered to be 100% complete?
A. The 1049 stars known to be within 20 parsecs.
B. The 9110 brightest known stars in the sky.
C. The 5367 known clusters in the Milky Way.
D. The 20 fastest known (radial velocity) stars.
E. No sample could ever be 100% complete.

The intended answers are: 1-D; 2-E; 3-C; 4-(F = ma); 5-B; 6-B.

Since blackbody radiation is typically discussed in the context of stellar photospheres, question (1) probes for a comprehension of the general properties in a familiar local situation, but one that is very different from stars. If Wien's law is understood, then it will be apparent that cool objects must radiate primarily in the far infrared.

Question (2) probes the understanding of atomic spectra. While some specific knowledge is needed to answer both of those questions, that is not an unreasonable expectation.
Question (3) is more difficult, requiring some analytical thought. Since bright point sources of light are not commonly found in daily life, students cannot appeal to experience for an answer. Many students have probably never noticed that shadows rarely have sharp edges!
Question (4) may actually help students to put Newton's second law into context, recognizing that the *difficulty* of pushing something is just a measure of how much force is needed to achieve a given acceleration.
Kepler's second law is generally discussed in terms of its effect on the *speed* of a planet moving on an orbit. Question (5) puts the law into the less familiar context of the time required to move over an arc of an ellipse.
Question (6) is the most challenging since it requires not only comprehension of the concept of statistical completeness and incompleteness but also some knowledge of how such things as the nearest stars are located, and why selection effects disturb surveys like that.

2.3. Application/Apply

An elementary course in astronomy is usually not sufficient to permit much application of the concepts that are learned and even comprehended. However, a rudimentary form of application can be achieved by means of a few strategies. One strategy makes use of images (objects, graphs, diagrams) projected during the examination, with the questions directed at each one. A modicum of coordination is required to carry this out within the limitations of time that usually prevail. There is an almost irresistible temptation to use this format for identification of types of objects like globular clusters, planetary nebulae, etc. and that is a legitimate use of this format under the category of KNOWLEDGE.

However, it takes on the category of APPLICATION if, for example, stellar spectra are projected and the question asks to identify which star is the hottest; which the coolest; which is a binary; which is the fastest; etc.

Light curves of eclipsing binaries can provide more applications by asking which one has only partial eclipses, which one has two stars with nearly the same temperature, which one has two stars with the biggest difference of surface temperature, or which one probably has the shortest orbital period (or which one shows evidence of significant tides); etc.

Color-magnitude diagrams for clusters provide a format for asking which cluster is the youngest; which is the oldest; etc.

Application of the Stefan-Boltzmann law of black-body radiation can be had by constructing an *idealized* H-R diagram with stars located at temperature coordinates of 3000 K, 6000 K, and 12 000 K along the main sequence, and designated as (A), (B), and (C), respectively. The luminosities of these stars are 10^{-2}, 1, 10^{3}, respectively, in units of the solar luminosity. Star (D) is then located at a temperature coordinate of 3000 K, with a luminosity of 10^{2} in solar units (simulating a cool giant), and star (E) is placed at temperature coordinate 12000 K with a luminosity of 10^{-2} in solar units (simulating a white dwarf). The line of questioning can then proceed as follows.

1. By what factor is the radius of star (D) larger than that of star (A)?
   A. 10
   B. 100
   C. 1000
   D. 3
   E. 30

2. By what factor is the radius of star (D) larger than that of star (B)?
   A. 10
   B. 20
   C. 40
   D. 100
   E. Not possible to compute from information given.

3. Identify the star that must have the largest radius.

4. An observer on Mars (1.5 A.U. from the Sun) would find the parallax of the nearest star to be
   A. 50% larger than it is measured from Earth.
   B. 50% smaller than it is measured from Earth.
   C. the same as it is measured from Earth.
   D. unmeasurable.
   E. variable.

The intended answers are: 1-B; 2-C; 3-D; 4-A.

Whether or not the first 3 of the above questions are truly APPLICATION depends on what the students have been asked to do with H-R diagrams during class or in homework assignments. By making the arithmetic trivially simple, the questions probe the ability to apply the blackbody law to stars, using nothing but scaling relations. The more able students recognize that question (3) is a contest between stars (C) and (D). Since the surface temperature of star (C) is 4 times larger than that of star (D), its surface radiates 256 times as much energy per unit area. If the two stars were of the same radius, that would be the ratio of their luminosities. Since the actual ratio is just a factor of 10, star (D) must have a larger radius than star (C) (specifically about 5 times larger). Clearly, question (3) crosses from pure APPLICATION to some ANALYSIS. Other variations of this question result by substituting 9000 K, 15 000 K, or 30 000 K for 12 000 K and by taking some liberties with the luminosities.

Question (4) is an application of the concept. One might argue that it is COMPREHENSION.

Another strategy for APPLICATION is to present an actual model for something like the interior of the Sun or of a star, or for the structure and dynamics of the Milky Way galaxy, etc. The line of questioning then involves a direct interpretation of the model. Table 1 is an example using a solar interior model.

Use Table 1. or Figure 1, to answer the next 6 questions.

Table 1. A model for the interior of the Sun.

Fraction of the radius	Temperature (K)	Fraction of the mass	Fraction of the luminosity
0.00	15 500 000	0.000	0.000
0.04	15 000 000	0.008	0.080
0.10	13 000 000	0.070	0.420
0.20	9 500 000	0.350	0.950
0.30	6 700 000	0.640	0.998
0.40	4 800 000	0.850	1.000
0.50	3 400 000	0.940	1.000
0.60	2 200 000	0.982	1.000
0.70	1 200 000	0.994	1.000
0.80	700 000	0.998	1.000
0.90	310 000	0.999	1.000
1.00	6000	1.000	1.000

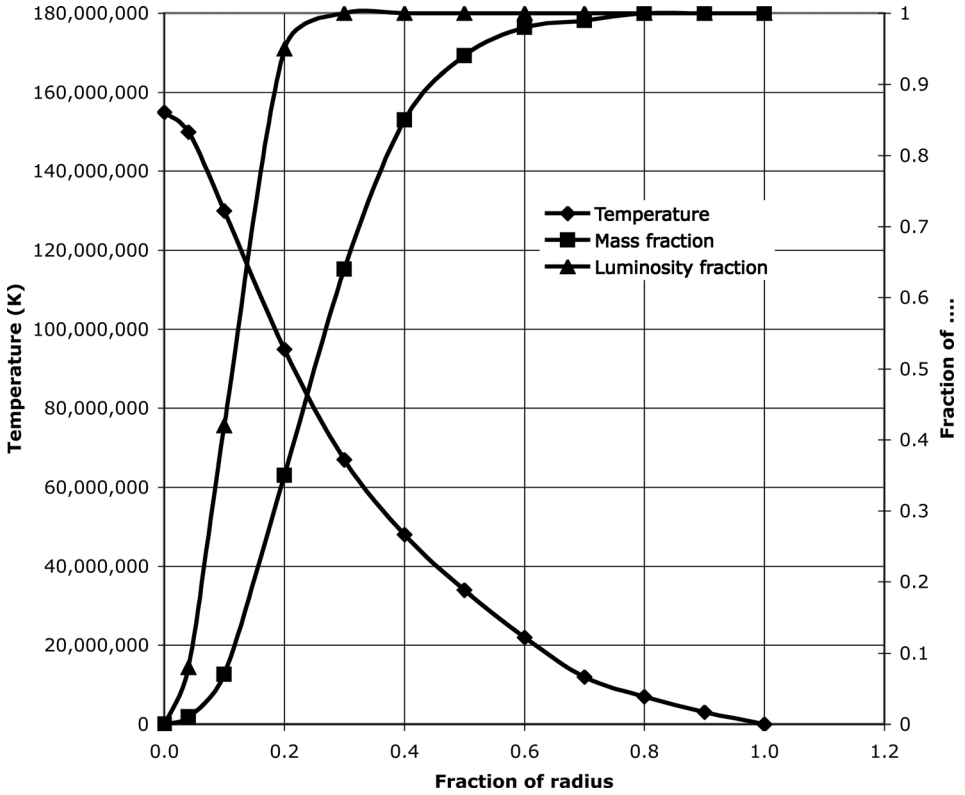

Figure 1. A stellar model for the Sun

1. Which of the columns lists actual measured data?
   A. Fraction of the radius.
   B. Temperature.
   C. Fraction of the mass.
   D. Fraction of the luminosity.
   E. None.

2. What fraction of the mass of the Sun generates 95% of the luminosity of the Sun?
   A. 5%
   B. 15%
   C. 25%
   D. 35%
   E. 50%

3. Approximately what fraction of the VOLUME of the Sun generates 95% of the solar luminosity?
   A. 1%
   B. 5%
   C. 10%
   D. 20%
   E. 30%

4. Approximately how much more *dense* is the central core of the Sun (innermost 4%) than is the entire Sun as a whole?
   A. 8 times.
   B. 50 times.
   C. 125 times.
   D. 500 times.
   E. 1000 times.

5. At what temperature inside the Sun does the model indicate that ALL thermonuclear fusion of hydrogen into helium has stopped?
   A. 15 000 000 K
   B. 13 000 000 K
   C. 9 500 000 K

D. 6 700 000 K
E. 4 800 000 K

6. At approximately what fraction of the radius of the Sun is the temperature GRADIENT largest?
   A. 0.02
   B. 0.07
   C. 0.15
   D. 0.25
   E. 0.35

The intended answers are: 1-E; 2-D; 3-A; 4-C; 5-E; 6-C.

Question (2) requires no more than the ability to interpret the actual meaning of the tabular values (i.e., read the table properly).
Question (3) requires some simple calculation and knowledge of how volume depends on radius for a sphere.
Question (4) may be the most difficult for non-mathematical students. A calculator might be helpful to insure correct arithmetic, but the real challenge is to decide how to utilize the numbers in the table (i.e., APPLICATION).
Question (5) is a particularly good example of APPLICATION of knowledge since it asks for something that is not explicitly tabulated.
Question (6) is also a good example of APPLICATION since it requires that the student understands exactly what a temperature "gradient" means and how that manifests inside a star. Again, a calculator would help to eliminate mistakes in arithmetic.

One of the most charming strategies for testing APPLICATION is to simulate a "Walk Under the Night Sky," whereby the student taking the examination is to imagine that he or she is walking with a friend at night, and the friend asks probing questions about *what is actually seen*. This is as much a learning experience as a testing experience since the student must connect abstract astronomical knowledge with things that can be seen directly with human eyes. It is also directed at playing down the importance of knowing names of stars and of constellation patterns, and substituting meaningful physical knowledge instead. The first author often suggested to students that they might prepare for this part of the examination by actually taking a friend (not a member of the class) for such a walk, and responding to his or her questions. They are cautioned that their friend will undoubtedly ask questions that are far more difficult than any that we would ask them!

2.3.1. Application Example: A Walk Under the Night Sky

Now that you have completed a course in modern astronomy you could take a friend outside on some clear, dark night and explain to him (or her) some remarkable things about what is actually seen. For example, you could point out that stars have different colors and show your friend an example of an orange colored star and a blue colored star (even though you do not know their names).

1. You would tell your friend that the different colors are caused by
   A. different chemical compositions.
   B. Doppler shifts caused by their motions.
   C. different surface temperatures.
   D. different luminosities.
   E. absorption by interstellar dust.
2. You would then point out that the orange colored stars must be
   A. the coolest ones.
   B. the hottest ones.
   C. the most luminous ones.
   D. the densest ones.
   E. the smallest ones.
3. Your friend asks how you can tell which objects are stars and which are planets. You reply that it is not apparent from just looking, but planets reveal themselves by
   A. slightly changing color each night.
   B. slightly changing position among the stars each night.
   C. slightly changing brightness each night.

D. setting after the Sun in the west each night.
E. Rising before the Sun in the east each morning.

4. Your friend asks if the stars shine by reflecting sunlight, like our moon does. You reply that they emit light of their own, and we know that because
   A. stars appear to have different colors.
   B. all stars show emission line spectra.
   C. stars twinkle, but the Sun gives steady light.
   D. stars are not large enough to reflect any light.
   E. the Sun does not shine at night.

5. Your friend asks, "Are the brightest stars the ones that are closest to us?" You reply that they are not necessarily the closest because
   A. brightness is not related to distance.
   B. big stars always appear brighter than small stars.
   C. hot stars always appear brighter than cool stars
   D. binaries emit twice as much light as single stars.
   E. luminosities differ by more than brightness.

6. Your friend asks, "Of all the stars we can see in the sky tonight, which is the largest one?" You search all over the sky and then point to one that is
   A. bright and orange.
   B. Bright and blue.
   C. Twinkling most rapidly.
   D. Brighter than all other visible stars.
   E. Moving faster than the other stars.

7. Your friend asks if the stars will shine forever, and you tell him (her) that every star must die someday because
   A. Eventually everything must die.
   B. All hot things eventually cool down.
   C. Stars consume a finite supply of hydrogen fuel.
   D. The universe is expanding.
   E. They fall into a black hole in the center of our Milky Way galaxy.

8. Your friend asks, "What happens to stars when they die?" You reply that most of them become
   A. black holes.
   B. brown dwarfs.
   C. red dwarfs.
   D. white dwarfs.
   E. neutron stars.

9. Your friend asks, "Can we see any dead stars in the sky tonight?" You reply that none are visible because
   A. no stars have died yet in our galaxy.
   B. dead stars do not radiate any light.
   C. it is impossible to tell which ones are dead.
   D. dead stars are only found in clusters.
   E. dead stars have very low luminosity.

10. You then show your friend that we are inside a great stellar system shaped like a disk by pointing out
    A. the line of the ecliptic.
    B. the bright blue stars.
    C. the horizon.
    D. the Milky Way.
    E. the North Star.

11. Your friend then asks, "If all the stars are in the galactic disk, then why do we see most of them all around us, far from the Milky Way?" You think for a moment until you realize that this is evidence that
    A. most stars are not really in our galaxy.
    B. most visible stars must be very near to us.
    C. interstellar dust must all be very far from us.

D. the Sun is located in an unusual place.
E. we are very near the center of our own galaxy.

12. Your friend asks, "If the universe is expanding, are all the stars in the night sky getting farther away from us?" You inform him (her) that such a notion is incorrect because
    A. only the remote regions of the universe expand.
    B. stars are very tightly bound to their host galaxy and some nearby galaxies (and their stars) are moving *toward* us.
    C. our Sun prevents the nearby stars from expanding.
    D. the expansion of the universe has stopped by now.
    E. it is the scale of spacetime that is expanding and not galaxies moving *through* space.

The intended answers are: 1-C; 2-A; 3-B; 4-A; 5-E; 6-A; 7-C; 8-D; 9-E; 10-D; 11-8; 12-E.

As you walk home at the end of a delightful evening of star gazing, your friend asks you the name of a bright orange colored star that is nearly overhead. You do not know its name, or even if it has one; but you recognize that it must be a very distant, evolved red giant. You have come to know the stars in a more meaningful way than by the arbitrary names given to them by some people who lived long ago and who knew nothing about what the stars really are.

2.4. Analysis/Analyze

We have treated APPLICATION as an ability to utilize knowledge and COMPREHENSION to account for real situations, and to recognize things in context. ANALYSIS is a more sophisticated version of APPLICATION, requiring that knowledge and comprehension be used to deduce the circumstances of an unfamiliar situation. If COMPREHENSION is the hallmark of good elementary science courses, then ANALYSIS is the central theme of advanced courses. In such courses ANALYSIS is taught and tested for by means of contrived problems. A similar strategy can be used in the General Education science course, albeit with decreased emphasis upon the use of formal mathematical operations. It is not inappropriate, however, to expect students in such courses to be able to carry through the arithmetic associated with the use of scaling laws. The following questions are designed to elucidate analytical thought in both qualitative and quantitative motifs.

1. If a simple convex lens has a focal length of 10 cm when it is illuminated with RED light, then its focal length when illuminated with BLUE light will be
   A. somewhat shorter than 10 cm.
   B. somewhat longer than 10 cm.
   C. precisely 10 cm.
   D. blue light cannot be brought to a focus.
   E. impossible to make such a prediction.

2. If a simple convex lens were immersed into a tank of water, its focal length in any color would
   A. remain unchanged.
   B. becomes much longer.
   C. becomes much shorter.
   D. not be measurable under water.
   E. no longer exist (cannot focus light under water).

3. Suppose that the eccentricity of the orbit of our moon were 0.8 instead of its actual value. How much larger would it appear to us when it was at perigee, than at apogee?
   A. 9 times larger.
   B. 1.8 times larger.
   C. 2.8 times larger.
   D. 4 times larger.
   E. It would appear the same size.

4. Astronauts working on a permanent base on the Moon would notice that the Earth
   A. rises in the east and sets in the west.
   B. rises in the west and sets in the east.
   C. never rises, sets, or moves.

D. is above the horizon only for 2 weeks each month.
E. rises and sets once each sidereal lunar month.

5. If, instead of the Earth, the Sun had a binary companion star with the same mass as the Sun in our orbit, the orbital period of that binary would be
   A. 2 years.
   B. 1.4 years.
   C. 0.7 years.
   D. 0.5 years.
   E. 1 year (same as the Earth's current period).

6. The orbital velocity of the Earth is about 30 km/s, and the velocity of light is nearly 300 000 km/s. Our motion causes the wavelength of hydrogen-alpha photons (6563 Å) from stars on the ecliptic to change by about
   A. 1.5 Å.
   B. 3.5 Å.
   C. 0.6 Å.
   D. 0.3 Å.
   E. 0.1 Å.

7. The Sun is like a black body radiating at a temperature of about 6000 K. The brightest wavelength that the Sun emits is about 5000 Å. What is the brightest emitted wavelength from a star whose temperature is 30 000 K?
   A. 100 Å.
   B. 300 Å.
   C. 500 Å.
   D. 1000 Å.
   E. 3000 Å.

The intended answers are: 1-A; 2-B; 3-A; 4-C; 5-C; 6-C; 7-D.

Questions (1) and (2) require some analytical thinking without any mathematical structure. Question (2) is far the more difficult, testing the comprehension of refraction as depending upon the ratio of the indices of the two media, and how lenses focus light.

Question (3) requires knowing that the escape velocity is the square root of 2 times larger than the circular velocity. It tests analytical skill to recognize that the Earth has the circular velocity, so it is necessary to remove 100% of that to send a probe to the Sun, but only to increase it by about 41% to leave the solar system.

Question (4) requires knowing the definition of the eccentricity of an ellipse, and then a geometrical analysis to find the ratio of the distances of the two extremes from a single focus.

Question (5) requires that the equation for the circular velocity be known, but we provide such equations on a supplementary sheet to discourage memorization and encourage comprehension. This question is designed to give an appreciation for the use of such equations as scaling laws.

Question (6) tests the ability to use knowledge in a new way. Students should know that the Moon rotates synchronously with its sidereal period, keeping one hemisphere facing us. It takes some imagination to see that the Earth is therefore a *synchronous satellite* from the point of view of an observer on the Moon.

Question (7) requires that Kepler's III law be used as a scaling law, but including the dependence upon mass that Kepler himself did not realize.

Question (8) is a straightforward application of the Doppler equation.

Question (9) requires that Wien's radiation law be treated as a scaling law.

2.5. Synthesis/Create

The taxonomic class that we designate as APPLICATION involves the utilization of knowledge and comprehension to provide explanations. The class called ANALYSIS is the more sophisticated version of application that make predictions about unfamiliar circumstances by making use of knowledge and comprehension. By the taxonomic class called SYNTHESIS, we mean either of the previous two classes (APPLICATION;

ANALYSIS) in which two or more *different* concepts must be put together to achieve the desired result of explanation or prediction. This kind of thinking can be challenging to even the most adroit students since they are generally taught to think in serial fashion. Thus, a good examination in a general education science course should not contain very many questions of this type, since they test for skills that are not usually present in most of the students. It is unreasonable to expect that this level of sophisticated thinking can result from the brief and superficial treatment of science that most general education courses are obliged to provide. However, it is precisely questions of this type that provide the discrimination for the highest grades achieved on any one examination.

The strategy for creating questions that require SYNTHESIS is exactly the same as that for APPLICATION and ANALYSIS, with the exception that two different concepts are superimposed into the same situation. One very good way to achieve this is to consider how various phenomena would appear as observed from places other than our normal vantage point on the Earth or in the galaxy. Some of the following examples make use of that strategy.

1. The retrograde motion of Mars occurs at intervals of
   A. one Martian synodic period.
   B. one Martian sidereal period.
   C. one Earth year.
   D. one Earth month.
   E. unpredictable duration.

2. Astronomers on Earth find that the distance to the nearest star (alpha Centauri) is 1.30 parsecs. If there were astronomers on Jupiter (5 A.U. from the Sun) they would find the distance to this star to be
   A. 1.30 parsecs.
   B. 6.50 parsecs.
   C. 0.26 parsecs.
   D. 13.0 parsecs.
   E. 0.13 parsecs.

3. An astronomer on Mars (1.5 A.U. from the Sun) would find that the effect of stellar aberration there is
   A. non-existent.
   B. larger than it is measured from Earth.
   C. the same as it is measured from Earth.
   D. smaller than it is measured from Earth.
   E. reversed from the way it appears from Earth.

4. The Earth orbits the Sun with a velocity of about 30 km/s. The velocity of light is about 300 000 km/s. Astronomers who observe the 3 K cosmic background radiation in the direction of the orbital motion of the Earth will find it to be
   A. cooler by 0.0003 K.
   B. hotter by 0.0003 K.
   C. cooler by 0.0001 K.
   D. hotter by 0.0001 K.
   E. unchanged since it comes from the universe.

5. Voyager pictures showed that the Great Red spot on Jupiter rotates counterclockwise. Since the spot is in the southern hemisphere of Jupiter, this observation shows that
   A. material is flowing into the Great Red Spot.
   B. material is flowing out of the Great Red Spot.
   C. material in the spot does not flow in or out.
   D. the Great Red Spot must be turbulent.
   E. the Great Red Spot must be a solid object.

The intended answers are: 1-A; 2-C; 3-D; 4-B; 5-B.

Question (1) links the explanation for retrograde motion with the concept of the synodic period.

Question (2) probes the concept of parallax in a broader context, illustrating its generality, and that it is not a phenomenon of the Earth. The student must connect the concept of parallax and that of distance, recognizing that although the distance to the nearest star is the same from Mars as it is from Earth, the parallax is not the same since it is a *relative* measurement (i.e., the ratio of the distance of the planet from the Sun to the distance of the star from the Sun). The question can be made more difficult by asking about the proper motion of a star. In an

absolute sense, the angular velocity of a star seen from Mars is the same as when it is seen from Earth. However, since proper motion is expressed as arc seconds per *year,* it will be found to be smaller by an observer on Mars because the year is longer.

Question (3) addresses this issue more directly, requiring the student to recognize that the parsec is a *relative* unit of distance, and not an absolute unit. Since the orbit of Jupiter is 5 times larger than the orbit of the Earth, the (Jupiter) parsec is 5 times larger than that defined by the orbit of the Earth. Thus, the distance to the nearest star is a smaller number of (Jupiter) parsecs. Greater emphasis on SYNTHESIS can be had with this question by stating it in terms of distance measured in light-years. The distance to the nearest star is about 4 light-years as measured from the Earth. However, as a consequence of Kepler's III law, the year for Jupiter is longer than the year for the Earth. Since the velocity of light is an absolute constant, the distance to the nearest star in (Jupiter) light-years is smaller than from the Earth by just the ratio of the orbital periods of the two planets. Since the size of the orbit of Jupiter is specified in the question, the student is expected to use Kepler's III law to estimate the orbital period. Jupiter's orbital period is nearly 12 years, so the distance to the nearest star would appear from Jupiter to be about 1/3 of a light-year.

Question (4) is a classic example of SYNTHESIS in which the phenomenon of aberration is linked with the physics of orbital motion. The student must recognize that aberration is dependent entirely upon the velocity of an observer relative to the velocity of light, *and* that the velocity of a planet that is farther from the Sun than the Earth must be smaller than the velocity of the Earth. Inner planets can be substituted for outer planets for variety, and the question can be made more difficult by requesting a quantitative answer.

Question (5) is an elegant illustration of SYNTHESIS, but it may be too difficult for students in the typical introductory course. Wien's radiation law for blackbodies is linked to the Doppler effect in this question. Some familiarity with differential calculus and how it is applied is needed to show that Wien's law leads to a fractional change of temperature that is equal to the fractional change of wavelength induced by the Doppler effect. The *fractional* change of wavelength is 10^{-4} and so the *actual* change of temperature observed is 3×10^{-4} K. The better students may not be able to see all the way to a final solution, but they should recognize that the effect of the motion must be to increase the apparent temperature in the direction of motion, and thus they have to decide only between choices (B) and (D). Instead of the orbital motion of the Earth about the Sun, one could specify the more important motion of the Sun about the galaxy, approximating it as 300 km/s for the purposes of the question. This brings the variation of the observed temperature to millikelvins, which is actually observed.

Question (6) links the Coriolis effect and the weather, including cyclonic and anti-cyclonic flows. Furthermore, it requires a clear understanding of the hemispheric inversion on a rotating sphere.

In a "Walk Under the Night Sky" the interlocutor could raise the hypothetical issue of what the night sky might look like if the Sun were located deep inside of a globular cluster. A considerable amount of SYNTHESIS is required to deal with this novelty. The absence of a "Milky Way," and of local interstellar dust, and the preponderance of densely packed red giants distributed over the entire sky would provide much food for thought.

2.6. Evaluation/Evaluate

A detailed description of the meaning of EVALUATION in the context of General Education science was given near the end of Section 1.3 along with an example of one type of question designed to test for one facet of that skill. Often, but not always, EVALUATION requires that a comparison be made between two or more alternatives, and that a judgment be rendered. The concern of educators with the issue of EVALUATION is often focused upon the ability to distinguish between science and pseudoscience, and with various religious beliefs that are inconsistent with scientific viewpoints. Equally important, however, for citizens who vote in a technological society is an ability to evaluate two conflicting *scientific* arguments. Performance in this latter capacity generally hinges upon the ability to assess the reliability of data, the validity of assumptions, and the assertion of model dependent facts. Testing strategies should therefore aim at those qualities.

1. Galileo's physics of motion described the path of a projectile as a parabola. What assumption produced that result?
   A. Projectiles do not accelerate while in flight.
   B. Neglect the effect of air resistance.
   C. Gravity is constant above the surface of Earth.
   D. Projectiles follow the curvature of the Earth.
   E. No assumption is needed for that result.

2. A popular book claimed that a near alignment of planets would cause tidal disasters on the Earth. That claim was absurd because
   A. planets can never really be aligned.
   B. the Moon blocks the effects of such planets.
   C. there is no gravity from other planets.
   D. tidal forces from planets are infinitesimal.
   E. the Earth is solid and resists tidal stresses.

3. If the universe is 10×10^9 years old and stars convert hydrogen into helium, why is there any hydrogen still remaining?
   A. The universe is only 6000 years old.
   B. There is almost no hydrogen still remaining.
   C. Helium is reconverted to hydrogen when stars die.
   D. Most hydrogen remains in interstellar gas clouds.
   E. Stars only convert 10% of their hydrogen into helium.

4. If stars have been EVOLVING by converting hydrogen into helium, why do they all have the same chemical composition?
   A. The universe is only 6000 years old.
   B. Conversion occurs only deep inside the cores.
   C. Interstellar hydrogen replenishes the supply.
   D. Stars do not appear to have the same Composition.
   E. Convection keeps stars uniformly mixed.

5. If the universe is really expanding, why do galaxies not appear to get smaller every year?
   A. The universe is not expanding.
   B. Light does not travel in exactly straight lines.
   C. Galaxies are not all the same size.
   D. The variation is too small to measure.
   E. Sizes of galaxies expand at the same rate.

(See also the model-dependent/independent question near the end of Section 1.3.)

The intended answers are: 1-C; 2-D; 3-E; 4-B; 5-D

Question (1) probes the subtleties of selection effects, and it can only be asked near the end of an astronomy course after the statistics of stars, binaries, and galaxies have been discussed. However, similar questions can be constructed within a more restricted topic area.

Question (2) is a good example of the need to be sensitive to the effects of a subtle assumption. It is distressing to see how many physics text books give a formal derivation of the path of a Galilean projectile as a parabola, without ever pointing out the assumption upon which it is based (and the fact that the assumption is false, and is in direct violation of the Newtonian theory of gravitation)! Many students who have taken an undergraduate physics course are convinced that the path of a (local) projectile is a parabola, and that it is a mathematically proven fact. Therefore, question (2) would be a challenge even for some physics students.

Question (3) makes reference to the pseudoscience book, "The Jupiter Effect," that preyed upon and exploited public ignorance; and questions (4), (5), and (6) are based directly upon some published arguments used by Biblical fundamentalist creationists to argue that the Earth (and everything else) is no older than 6000 years. That assertion is given as the first choice for an answer in each case, to provide the format for EVALUATION since it is the correct choice for a young-Earth creationist. Most students will eliminate the naive fundamentalist view, but the better part of critical thinking is to identify a good argument, and not merely to reject a poor one.

Since the model-dependent vs. mode1-independent format was given earlier, it was not repeated here. However, this is the most important aspect of the skill of EVALUATION. Although practicing scientists are keenly aware of the distinction, both in general and in their own work, they are often not cautious about blending the two in the heat of a discussion. Their colleagues instinctively sort out such blends, largely because they are aware of what things are models in their discipline. However, the lay public lacks such awareness and is easily deceived by what sounds like a cogent argument. Students in a one-semester general education science course cannot hope to gain such a sophisticated discriminatory sense, but with some practice, it is possible to discern the general traits of model dependent assertions. If nothing else, it is a mark of some distinction to be aware of the difference and to attempt to seek it out when EVALUATING a scientific presentation.

The remainder of the paper consists of appendices that readers may find useful: Appendix A. Architecture of Questions; Appendix B. Theory of Foils; Appendix C. Statistics of the Responses; Appendix D. Grades and Scale Normalization; Appendix E. Combining Examination Grades for a Final Grade; and, Appendix F. Significant Figures.

Acknowledgements

This article was entirely conceived of and written by Dr. Art Young who passed away on February 7, 2012. As Dr. Young's coauthor, Dr. Shawl's role was one of encouraging Dr. Young to publish the paper, which had sat idle for at least a decade, and then spending considerable time and effort in the actual preparation for publication while Dr. Young's health was declining and after his death. Changes to Dr. Young's original manuscript were made in response to reviewer comments.

The authors would like to thank the reviewers for their cogent comments on a less than perfect manuscript, and for mentioning some specific useful references. The reviewers' comments materially improved this paper, for which the authors are grateful.

Appendix A. Architecture of Questions

In the previous section, we considered the construction of questions designed to test for specific attributes under the general topic of the course. There are some principles of design that transcend the specific topics, and we now take up the general design of all questions regardless of the attributes that are being tested. The first of these general principles concerns the construction of the foils, i.e., those incorrect answers that are planted intentionally to conceal the correct answer. Then we look at the statistics of responses, grades and normalization, combining grades in reaching a final grade, and significant figures.

Appendix B: Theory of Foils

Creating the foils is probably the most challenging and time consuming aspect of the construction of multiple choice examination questions (Haladyna and Downing 1989). The basic design concept behind the foils is that they should serve a useful diagnostic function on an individual basis for students, and on a collective basis for the quality of the questions and of the examination as a whole. The foils will, ideally, also provide the instructor with information to help improve the teaching of concepts with which students have difficulties.

There are some general principles that should first be enumerated prior to examining the diagnostic functions. The foils should conform as much as possible to the following design specifications.

1. There is no set "best" number of foils but 3–5 is typical (Delgado and Prioto 1998). The material and the ability to produce valid foils is the best criterion to use.
2. The foils should always be reasonable choices, or at least appear to be reasonable, to prevent simple elimination by students who do not really comprehend the question.
3. The better students should be able to eliminate 3 of the 5 choices, but need to deliberate carefully with the remaining 2 choices.
4. The options, "none of the above," and "all of the above" should be used sparingly. Appropriate and inappropriate uses of these options will be discussed later.
5. All of the responses, including the correct one, should be as short as it is possible to make them without compromising clarity or good grammar. If possible, they should be contained on a single line; words in common to all can be placed as part of the stated question.
6. Complete sentences and proper grammar should always be employed.

The diagnostic properties of the foils reside in their ability to aid the process of analyzing poor performance by a student on the examination as a whole. If the foils are created with some care, it should be possible to distinguish between students who have not been diligent in their studies, and those who have been diligent but who need additional instruction for clarification. In more restricted topic areas such as orbit mechanics, or light and radiation theory, etc., it should be possible to identify a specific type of misconception that persists with an individual student who is otherwise doing well.

Whenever a commonly held misconception is known, it should be incorporated directly as one of the foils. Classic examples are: the changing seasons on the Earth is due to changes in the distance between the Earth and the Sun; artificial satellites remain in orbit because a centrifugal force balances the Earth's gravitational force; astronauts are weightless because they have escaped from the gravitation of the Earth. Other examples of common misconceptions are discussed in McDermott (1984) and Mestre (1991).

In cases where classic misconceptions are not prevalent, one of the foils should be a reasonable, yet definitely incorrect option. The remaining foils should depart more from being reasonable alternatives, while never becoming totally absurd. A student who consistently selects unreasonable options without any discernible pattern divulges a failure to grasp the essence of the subject for reasons that can be explored by an interview. Conversely, one who displays a pattern of usually selecting the correct response, but often selecting the reasonable but incorrect response, needs some specific guidance, but little more. Much can be learned from the latter student in an interview in which he or she is asked to try to reconstruct the thinking that led to the choices of the incorrect foils. If the foils are constructed strategically, one might uncover a specific misconception, such as failure to appreciate the fact that perfect blackbodies are perfect radiators as well as perfect absorbers. Another trait that can be discerned from the patterns of answers to properly designed questions is that of rote memorization of information with little or no comprehension.

When the option *"none or the above"* is used as a foil, it often means that the examiner was unable to construct enough substantive foils and has opted to use this device for no other purpose, either didactic or analytic. There is some merit to that if the device is to be used as a correct choice in some questions, for otherwise it will become apparent that whenever it appears it *is* the correct choice! When it is the correct choice, it has little didactic value unless its purpose is to call attention to several notably invalid ways to explain something. Otherwise, the student may be left with four incorrect explanations, which he certifies by selecting *"none of the above",* and is then left to wonder if there are *any* correct explanations. This device can be described as a "non-answer, answer" and as such it has no counterpart in essay type examinations. Students perceive it as intentional obfuscation, and well it may be in some cases, thereby making its use of questionable value for evaluation of knowledge or comprehension.

The converse device, *"all of the above,"* used as an affirming mechanism is less objectionable and may have some genuine didactic merit. This, too, should be used sparingly, and should be incorporated into some questions as a genuine foil to avoid a default condition. The device has the didactic merit of illustrating situations in which more than one factor is simultaneously present, or some similarity exists that might not have been apparent to students until asked in this particular manner. The following questions illustrate such uses of the device.

1. Why did astronomers fail to detect stellar parallax for centuries after Copernicus suggested looking for it as a test for heliocentrism?
   A. Stars are much more distant than was thought.
   B. Angle measuring devices were very crude.
   C. No good telescopes were available.
   D. Nobody knew which stars were actually nearest.
   E. All of the above were reasons for the failure.

2. Taken as an entire class, the outer planets (those beyond Mars) differ from the inner planets in
   A. mass.
   B. radius.
   C. rotation.
   D. chemical composition.
   E. all of the above.

3. A rainbow is a good example of
   A. the spectrum of black-body radiation.
   B. refraction.
   C. total internal reflection.
   D. decomposed white light.
   E. all of the above.

The intended answers are: 1-E (although student might be able to argue A as the one best answer!); 2-E; 3-E.

Appendix C: Statistics of the Responses

Multiple choice examinations should always be machine-scored not only for the speed, accuracy, and convenience that is afforded, but also to obtain statistical data concerning the responses to each of the questions. Ambiguous questions, poor foils, or inadequate instruction can often be detected by a study of such statistics. This kind of analysis should be done every time that the questions are used, even if they have been in service for some time, but especially when new questions are introduced.

There is no single criterion for a good question. For example, one could have a good question to something for which the students have a significant misunderstanding or misconception. That not withstanding, generally such questions would have more than 50% of the class giving the correct response, and the bulk of the others selecting a foil that was either the most reasonable alternative, or a prevalent misconception. Understanding the choices made by students allows the instructor to mediate the problem. A significant literature exists on this subject.

Some of the more common signs of a problematical question are the following.

1. A very large fraction of the class selecting the correct response (95% or more).
2. A very small fraction of the class selecting the correct response (less than 15%).
3. A large fraction of the class (50% or more) selecting a particular foil.
4. Uniformly distributed responses across most or all of the foils; i.e., nearly 20% of the class for each foil.

Situation (1) usually means a very easy question, or one whose foils are weak and easily eliminated. A few such questions are acceptable, but they should be examined to see if a change in wording of the question, or in one or more of the foils, would be helpful to make the question useful for evaluation.

Situation (2) is ambiguous and requires further study. The question may be too difficult; or it may be worded so as not to be understood or ambiguous; or the instruction on that topic may have been inadequate. A study of the statistics of the responses to the foils may clarify that somewhat.

Situation (3) usually means that the "correct" response was poorly written, so a reasonable foil appears better. This might also mean an inherent ambiguity in the way in which the question is posed. Generally, it is immediately apparent when the question is reviewed in light of the foil that is heavily selected. This is a common occurrence since the person who constructs the questions has a clear idea of what was intended, and rarely interprets things in the manner that students do.

Situation (4) usually points to inadequate instruction about the topic so that all of the foils appear reasonable.

Item analysis should be an ongoing process when multiple choice examinations are utilized, and improvements are always possible. Multiple versions of the same question should be kept in the test bank, and their statistics compared between different classes. A randomization scheme should be used to insure that no choice-letter is favored for the best answer.

Appendix D: Grades and Scale Normalization

The multiple choice examination produces an objective and unambiguous score for the performance of each student. The number of correct responses and the fraction of the total number that are correct are measured. However, the evaluation of that score as a measure of achievement toward the goals of the course is subjective. In most institutions that evaluation is expressed as a letter grade on a scale of five (A–F), often further subdivided into a scale of 12 (adding plus and minus). A methodology for making that evaluation less subjective is highly desirable.

One school of thought holds that an absolute scale can be established that then converts the absolute score into an absolute performance. Generally such a scale is defined as 90% and above being graded as an "A," 80% to 89% as a "B," 70% to 79%' as a "C," 60% to 69% as a "D," and any score lower than 60% receives an "F" and is taken to be a failing performance. While those boundary values may be adjusted, the philosophy of this method assumes that all examinations are measuring instruments that are calibrated to a common set of units. Scientists know that such calibration can be achieved only if the instruments can all be set to measuring one common quantity before they are used to measure a variety of different quantities. In the case of examinations given by different instructors, or even those given by the same instructor at different times, such an objective calibration is not feasible. Therefore, using this methodology is the equivalent of attempting to compare measurements of lengths that are done in *centimeters* by one person, with those done in *inches* by another. Nobody would consciously make such a comparison because it is customary to label such measurements with units, such as *cm* and *in*. However, grades are tacitly labeled as if they were on the same scale (percentage; or letter equivalent) and so the invalid comparison is often done unwittingly.

As an example, we might imagine gymnastics instructors testing children for their ability to jump. Each instructor places one of his own hands upon a wall as high as he can reach, and scores the performances as percentages of that height reached by the top of the head of each jumping child. Upon comparing notes, the shortest instructor finds that he always has far more "A" students than does the tallest instructor!

Since absolute calibration of examinations is generally not possible, the best solution to the problem of evaluation and inter-comparison is relative scale normalization. This is accomplished by fitting a statistical probability distribution to the observed frequency distribution, in a process that is commonly known as *grading on a curve*. Most students do not know what this procedure means or how it is done, but they do seem to know that, like vitamin C, it is something that is good for them even if they have no concept of what it is!

Using this method, the measured score loses its typical meaning, and the performance of individual students is evaluated relative to the class as a whole. The assignment of performance grades is based upon a statistical model that, like any model, might not be a valid representation of the actual population. The standard model assumes that no unusual selection effects have operated to populate the class, and thus the performances of a large number of students will approach the random error distribution represented by the Gaussian distribution function.

The sample mean value then serves to normalize the scale of each examination, and the sample standard deviation measures the performance of the class on that examination. A grade of "C" is assigned to the mean score, and the other grades are assigned by an arbitrary criterion for how much departure from the mean score merits a different grade, either better or worse. This latter criterion is expressed in units of the standard deviation, and then it is called the Z-score. The assignment of a Z-score of +1 for a grade of "B," and +2 for a grade of "A" is as arbitrary and unjustified as assigning a percentage score of 80% for a grade of "B," and 90% for a grade of "A." A more rational, though still arbitrary, way of assigning the grades is to decide on the relative fraction of a class that should receive grades of "B" and "A" (and "D" and "F"). Such fractions then define the Z-scores at which the grades change and the correct values can be found in a table that is usually called the "error function" (or erf) in most statistics textbooks. Table 2 was constructed for a model in which 8% of the class is to receive a grade of "A," 16% "B," 53% "C," 15% "D," and 8% "F."

The sample mean is computed from

$$M = \frac{1}{N}\sum_{1}^{N} G_i.$$

The sample standard deviation is

$$S = \sqrt{\frac{\sum_{1}^{N} (G_i - M)^2}{N-1}}.$$

The symbology is: G_i = each score; M = mean of all scores; S = sample standard deviation; N = total number in sample.

The Z-score for each student is then given by the expression

$$Z_i = \frac{(G_i - M)}{S}.$$

Actual scores can never be truly Gaussian for many reasons, but when N is large (more than 100) the approximation is reasonable. Formal statistical tests, such as the chi-square statistic, can be used to determine the goodness of fit, but a simple plot of binned scores is sufficient to assure that there is no significant departure from a random distribution. The histogram shown below was made from actual scores on one examination for a class in astronomy. The plotted theoretical Gaussian probability distribution was calculated by using the sample mean and the

Table 2. Assigning letter grades based on Z-score.

Range of Z-scores	Letter grade	Fraction of class
$Z >= 1.4$	A	8%
$1.4 < Z >= 0.7$	B	16%
$0.7 > Z >= -0.8$	C	53%
$-0.8 > Z >= -1.4$	D	15%
$-1.4 < Z$	F	8%

sample standard deviation. That procedure is only an approximation for visual display. The correct procedure involves the use of a non-linear least squares fitting algorithm. When proper fitting is done, the derived values for the (model) mean and standard deviation should be used instead of the sample values to compute the Z-scores.

On the plotted graph, the percentages inscribed beneath each letter grade are the theoretical values that should result from the choice of Z-scores that define the vertical lines that are the cutoff values for each grade. The actual fractions in this particular class are: A 7%; B 10%; C 60%; D 15%; F 8%. One might decide, for example, that there are too many C grades, and not enough B grades resulting from the direct application of the formal criteria. A small adjustment of the Z-score criterion for the lower end of the B grades will correct that at once for this particular examination.

By using a method of this general sort, a table such as the one presented earlier can be published as part of the course description, so that students know how their relative performance will be graded. It might also be stated as policy that any small departures from the stated criteria will only be done if they affect the grades in a positive manner.

Appendix E: Combining Examination Grades for a Final Grade

The absolute percentage score method of assignment of grades leaves the dilemma of how to combine the results of all examinations since they are not on a common scale. Inevitably, one examination will be more difficult than another in the same sequence within a course, even when constructed by the same instructor. Some topics are more challenging than others for students in the introductory courses. Any direct arithmetic combination of examination scores is surely the equivalent of directly averaging distances that are given in centimeters with others that are given in inches.

The scale normalization that is achieved by fitting a Gaussian function to the distribution of scores eliminates this problem. In principle, it is possible to average the Z-scores that each student achieves on each examination for a final (mean Z-score) grade that can be assigned automatically from the table given previously. However, this is not the best way to proceed.

In moderate sized classes (more than 100, but fewer than 1000) the statistical. Fluctuations that occur naturally will disturb the tabular grade scale. Using the Z-score cutoff values given by the table, a particular examination may result in only 4% of the class obtaining a grade of "A," or some such anomaly. As indicated previously, this is readily corrected by making small adjustments in the actual Z-score cutoffs for each examination. Once that is done, however, the Z-scores are no longer normalized since they have slightly different meanings for each examination.

Therefore, the best procedure is to assign letter grades for each examination based upon the (modified) cutoffs for the Z-scores. The letter grades themselves correspond to numerical grade points that define an absolutely normalized scale. The grade point scale in Table 3 is (supposedly) universal among all American colleges and universities so as to permit transfers of students and other inter-comparisons.

Table 3. Grade points associated with each letter grade.

Grade	Grade points
A	4.0
A−	3.7
B+	3.3
B	3.0
B−	2.7
C+	2.3
C	2.0
C−	1.7
D+	1.3
D	1.0
D−	0.7
F	0.0

Thus, at the end of a term, a student will have a set of grade points corresponding to the various examinations, and the average of those values is a valid way to combine all of the performances for a final grade that is then taken from the table of grade points given above. If examinations are to be weighted differently, those weights can now be applied directly to the grade points as shown in the formula below where

$$G = \frac{\sum_{N} g_i w_i}{\sum_{N} w_i}.$$

G is the final grade (point), g_i is the individual grade points, and w_i is the weight. Spreadsheets on personal computers are ideally suited to keeping these kinds of records and carrying out these calculations.

Appendix F: Significant Figures

Mathematical constants, such as pi, have an absolute value that, if it is not rational, requires an infinite number of digits following a decimal point to express. The decision to truncate such a constant and express it (approximately) as 3.141593 does not necessarily imply a lack of knowledge of successive digits beyond the 6th place of decimals, but rather a lack of concern for them in a given circumstance.

Measured physical quantities, however, are generally not known to unlimited accuracy, even when they refer to something that does have a unique and absolute value (such as the mass of a hydrogen atom). The statement of any such value with a limited number of decimal places is not a truncation for convenience, but rather an expression of the limit of knowledge. The expressed numbers are referred to as *significant figures,* and they define, implicitly, the limit of current knowledge of the absolute value of the quantity. Although this concept is elementary, it is violated in common usage. It is not uncommon, for example, to find a container of milk that states the contents as 1 pint, or equivalently, 0.473176 liters! Apparently, only scientists seem to recognize the absurdity of this statement and not those who produce or consume milk. The quantity is given originally to one significant figure, although more could be justified since the amount of milk is probably controlled to about 1%, and thus it should be stated as 1.00 pint. (It is equally common for those who are not familiar with the concept of significant figures to omit trailing zeros even when they are significant and convey important information.) However, the conversion of units to liters utilizes the defining conversion constant without regard to significance, and therefore inadvertently improves the implied accuracy of the contents of the container to parts per million!

Another common abuse of significant figures occurs when quantities with limited accuracy are combined by arithmetic operations, and herein lays the relevance to grades. If, for example, the sides of a cube are measured and found to have the lengths, 1.2 cm, 2.6 cm, and 4.4 cm, the volume of the cube found in student laboratory notebooks is likely to be 13.728 cm. The lengths of the sides are only known to 2 significant figures, but somehow they yield a volume that is implied as being known to 5 significant figures! Since arithmetic cannot increase the amount of information that is known, this is a violation of the use of significant figures to represent accuracy.

However, if a single quantity, such as a length, is measured repeatedly to a fixed accuracy, the mean value of all of the measured values is better determined than is any one of the measurements. Thus, the mean value of such an array can justifiably be specified to one more significant figure than any of the measured values. The increased accuracy of the result does not come from the arithmetic, but rather from the *redundancy* of the measurements. It is this particular process that is generally misunderstood by all colleges and universities that specify grade point averages (GPA) of students to three significant figures.

From the table, it is apparent that grade points are quantities that are *defined* to have only two significant figures. A grade of B+, for example, has a grade point equivalent of 3.3 and not 3.30. However, it is the practice of colleges and universities to average all of the grade points from all of the courses that a student has taken, and then report the "grade point average" (GPA) with 3 significant figures. This is an invalid procedure since it assumes that the same quantity is being measured by each course (i.e., the student him/herself). However, since most courses are independent and uncoupled (except for a sequence of mathematics, science, or other discipline specialty courses) the quantity that is measured by each course is the performance of the student in that particular specialty, and not his/her innate ability. In general, course grades measure different quantities, so if they are averaged the result can only be designated as a "characteristic" quantity and not a "better estimate" of performance. For example, if one were to measure the length of each and every table in a furniture store to the nearest centimeter, and then average all of those values, it would be absurd to state that result to the nearest millimeter (one additional significant figure) since no single table has been measured to that accuracy, and the mean value does not

refer to any one table in the store (i.e., no redundancy). The average of the lengths of all of the tables in the store could be used only to "characterize" the lengths of tables that are sold in that store, and that quantity has little meaning without a specification of the *variance* of the lengths, which measures their distribution around such a mean value.

A more poignant example of how this misconception of significant figures affects grades is had from a true case story at an unnamed college. A particular scholarship was awarded to those applicants who had a GPA of 3.8 or greater. One applicant had a GPA (reported by the Registrar) of 3.79. The scholarship committee consisted of a physicist, a mathematician, and a philosopher. The physicist argued that the scholarship should be awarded to this applicant since grade points are measured quantities that have some uncertainty associated with them, and by an argument similar to the one above, three significant figures cannot be justified for that quantity. The mathematician, however, argued that awarding the scholarship is equivalent to stating that $3.79 = 3.8$ and that is certainly false. The philosopher argued that awarding the scholarship is the equivalent of altering a grade and that is immoral. A good student lost a scholarship to numerical ignorance!

References

Anderson, L. W., and Krathwohl, D. R. (Eds.). (2001), *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives: Complete edition*, New York: Longman.

Bardar, E. M., Prather, E. E., Brecher, K., and Slater, T. F. 2006, "Development and Validation of the Light and Spectroscopy Concept Inventory," *Astronomy Education Review*, 5, 103.

Bailey, J. M. 2007, "Development of a Concept Inventory to Assess Students' Understanding and Reasoning Difficulties about the Properties and Formation of Stars," *Astronomy Education Review*, 6, 133.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. (Eds.) 1956, Taxonomy of Educational Objectives: The Classification of Educational Goals. *Handbook 1: Cognitive Domain*, New York: David McKay.

Delgado, A. R., and Prieto, G. 1998, "Further Evidence Favoring Three-Option Items in Multiple-Choice Tests," *European Journal of Psychological Assessment*, 14, 197.

Haladyna, T. M., and Downing, S. M. 1989, "A Taxonomy of Multiple-Choice Item-Writing Rules," *Applied Measurement in Education*, 2, 37.

Haladyna, T. M., Downing, S. M., and Rodriguez, M. C. 2002, "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment," *Applied Measurement in Education*, 15, 309.

Hoffman, B. 1962, *The Tyranny of Testing*, New York: Crowell-Collier Press.

Hufnagel, B. 2001, "Development of the Astronomy Diagnostic Test," *Astronomy Education Review*, 1, 47.

Krathwohl, D. F. 2002, "A Revision of Bloom's Taxonomy: An overview," *Theory into Practice*, 41, 212.

McDermott, L. C. 1984, "Research on conceptual understanding in mechanics," *Physics Today*, 37, 24.

Mestre, J. P. 1991, "Learning and Instruction in Pre-College Physical Science," *Physics Today*, 44, 56.

Partridge, B., and Greenstein, G. 2003, "Goals for 'ASTRO 101': A Report on Workshops for Department Leaders," *Astronomy Education Review*, 2, 46.

Slater, T. F. 2008, "The First Big Wave of Astronomy education Research Dissertations and Some Directions for Future Research Efforts," *Astronomy Education Review*, 7, 1.

Wallace, C. S., Prather, E. E., and Duncan, D. K. 2011, "A Study of General Education Astronomy Students' Understandings of Cosmology. Part I. Development and Validation of Four Conceptual Cosmology Surveys," *Astronomy Education Review*, 10, 010106.

ÆR
010113-1–010113-27