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

Astronomy Education Review

2010, AER, 8, 010111-1, 10.3847/AER2009024

The Astronomy and Space Science Concept Inventory: Development and Validation of Assessment Instruments Aligned with the K–12 National Science Standards

Philip M. Sadler
Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts, 02138
Harold Coyle
Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts, 02138
Jaimie L. Miller
Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts, 02138
Nancy Cook-Smith
Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts, 02138
Mary Dussault
Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts, 02138
Roy R. Gould
Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts, 02138
Received: 03/10/09, Accepted: 11/18/09, Published: 01/11/10

Abstract

We report on the development of an item test bank and associated instruments based on those K–12 national standards which involve astronomy and space science. Utilizing hundreds of studies in the science education research literature on student misconceptions, we have constructed 211 unique items that measure the degree to which students abandon such ideas for accepted scientific views. Piloted nationally with 7599 students and their 88 teachers spanning grades 5–12, the items reveal a range of interesting results, particularly student difficulties in mastering the NRC Standards and AAAS Benchmarks. Teachers generally perform well on items covering the standards of the grade level at which they teach, exhibiting few misconceptions of their own. Teachers dramatically overestimate their students' performance, perhaps because they are unaware of their students' misconceptions. Examples are given showing how the developed instruments can be used to assess the effectiveness of instruction and to evaluate the impact of professional development activities for teachers.

1. INTRODUCTION

Black holes, frozen worlds, the "big bang," and supernovae: Astronomy and space science educators draw upon denizens of the celestial zoo more outlandish than the animals in any zoo on the Earth. As far as stories go, astronomers' tales are arguably stranger and more compelling than those of any other branch of science. Yet, there is more to astronomy and space science than incredible objects and extreme theories. The underlying concepts on which astronomy and space science are based are the traditional elements of physics, chemistry, and Earth science. Students' strong interest in astronomy can provide the motivation for learning these fundamental scientific principles. Understanding of the most basic astronomical structures is needed building blocks for learning about current research findings. Without being able to model the Earth-Moon-Sun system, the processes by which exoplanets are detected makes little sense. In our society, an understanding of basic astronomy and space science concepts by the average American contributes to informed decision making concerning future research and space exploration efforts (Kaufman 1997).

Two national efforts to characterize the knowledge required for a scientifically literate citizenry provide well-vetted listings of key fundamental concepts: The National Science Education Standards (National Research Council [NRC] 1996) and the American Association for the Advancement of Science (AAAS) Benchmarks (Project 2061 2001). These documents form the basis of curriculum and evaluation frameworks developed by all 50 U.S. states. Each document includes a substantial body of astronomical concepts (Adams and Slater 2000) at the primary, middle, and high school levels.

The astronomy education community responded to the challenge of the national standards by developing innovative ways to teach astronomy in the nation's schools. Soon after the first standards were released, new curricula [e.g., Project ARIES (see Note 1) and Astrobiology (see Note 2) and investigatory tools [e.g., MicroObservatory (see Note 3)] began vying for classroom time along with classroom visits by astronomers, both amateur and professional (e.g., the Astronomical Society of the Pacific's Project ASTRO). Only a few of these curricula and programs have undergone rigorous quantitative evaluations to measure their effectiveness (e.g., Sadler 1998; Ward, Sadler, and Shapiro 2007; Gould, Dussault, and Sadler 2007). Overall, there has been little in the way of evaluation of the combination of these efforts and the degree to which American students have increased their literacy in the domain of astronomy.

While the NRC and AAAS were specific concerning the particular science content knowledge required for astronomical literacy, they did not develop assessments that could measure such knowledge (Arnaudin and Mintzes 1985). The U.S. states have recently incorporated science tests into their assessment systems based on state standards. The National Assessment of Educational Progress administers science tests to a carefully selected sample of U.S. schools to monitor students' understanding of science, of which astronomy represents only a small part of these assessments. This paper reports on the development and validation of a bank of test items designed to assess understanding of all the astronomical concepts listed in the NRC standards and AAAS benchmarks. We seek to understand the extent to which students and their teachers have mastered these concepts and offer a method of measuring gains in astronomy content knowledge at the precollege level.

To accomplish this task first required the creation of a set of items that, as a whole, represent the entire body of the national standards relating to astronomy and space science content. These tests could then be given to a nationally representative sample of classrooms, in which both teachers and students could be included (Ausubel, Novak, and Hanesian 1978). Comparisons could then be made between the degrees of mastery of different standards in different grade bands. In particular, the strength of teacher knowledge for each of the standards could also be determined and related to student performance by classroom.

The assessment of science content is a complex issue and educators vary in their opinions about the best way, out of many assessment options, to measure student understanding. Tools such as portfolios (Slater 1997), clinical interviews (Duckworth 1987), "authentic" assessment (Kamen 1996), and concept mapping (Novak 1998) have grown out of the research on cognitive models of learning, attempting to characterize the path of an individual's conceptual change (Mintzes, Wandersee, and Novak 2005). Yet because of how they are formulated, these more qualitative methods are often much more expensive and complex to administer and score in order to collect usable data than more traditional standardized assessments (see Note 4). Over the past 20 years, a new type of assessment instrument has been developed based on qualitative cognitive research, but quantitative in format: The "distractor-driven" multiple-choice (DDMC) test. Research into the nature and potency of these tests shows great effectiveness in assessing the conceptual understanding of students (Halloun and Hestenes 1985).

2. BACKGROUND

The reader may find it useful to understand how these DDMC tests came into existence. Early cognitive research on children's ideas is usually traced to Piaget's traditional structured clinical interview (Piaget and Inhelder 1967). Piaget's work revealed that children build theories based on their interaction with the world (Turkle 2008), and these can be quite different from theories held by adults (Driver and Easley 1978) (see Note 5). While clinical interviews proved fruitful, they were time consuming. Prather (1985) identified a need for reliable diagnostic tests that could identify and classify students' conceptions. Open-ended written tests were developed as a way to draw from larger populations than interviews, followed by multiple-choice tests, the most well known being the Force Concept Inventory (Halloun and Hestenes 1985) in physics. Such written tests force a choice between a single correct answer and one or more misconceptions identified by researchers (Freyberg and Osborne 1985) (see Note 6). Such items can only be constructed if there is a cognitive model of the misconceptions that students hold on their way to a scientific understanding. Hence, knowledge

of what would make an attractive distractor drives the item creation. DDMC test items have the shortcoming that they limit incorrect responses to previously identified ideas (Finley 1986), demanding that items be based on the prior research concerning learners' misconceptions. However, well-crafted DDMC tests reasonably reproduce the results of lengthy and expensive interviews and can ascertain the conceptual frameworks of teachers as well as students (Gilbert 1977; Halloun and Hestenes 1985).

Recent research in psychometrics has explored the construction and application of collections of such items (see Note 7). Although DDMC tests remain uncommon, these recent studies suggest that the utility of instruments based on response patterns is growing. Well-constructed tests help science teachers have clearer ideas about the thinking of students. In addition, DDMC tests have the capability to identify both examples of student misconceptions as well as their prevalence/frequency within a population. Misconceptions, while they can change through instruction, appear to be quite stable in populations over long periods of time and appear to be similar across different cultures (see Note 8).

To those unaware that certain misconceptions are irresistibly attractive to learners, DDMC item answer choices look quite conventional. However, the statistical performance of these items is very different because a single wrong answer is chosen by the majority of students who answer the item incorrectly. Such a popular wrong answer, or distractor, is typically not found in a standardized test item. Psychometricians reject items with certain statistical profiles, particularly those for which moderately scoring students prefer a particular wrong answer with greater frequency than their lower performing classmates (Sadler 1998). This profile occurs when low-performing students guess at answers. The guessing strategy produces a probability of 0.20 correct on a five-item test. But when students hold a particular misconception, the popularity of a particular distractor will increase above 0.20 and the probability of choosing the correct answer can dip below 0.20. In other words, on a five-item test, students drawn to distractors may get no item correct while students who simply guess will, by chance, likely answer at least one item correctly. Therefore, including popular misconceptions makes DDMC test items more difficult and test developers advise against their use because they tend to "trap even very knowledgeable students (Nunnally 1964)."

Our belief is that this is the point of a good test item, i.e., does it distinguish between a student's preconceived ideas and those accepted by scientists? If the most prominent misconceptions are not included as distractors in test items, students may chose the correct answer via a process of elimination from a sea of "weak" distractors. Such items do little to inform teachers of their students' initial ideas when given as part of a pretest prior to instruction. More importantly, these items do not adequately measure the degree to which students have fully accepted the scientific concept when given on a post-test since students are not "tempted" by their misconceptions. By not presenting an item offering a choice that clearly reflects students' own thinking about a concept, students will often simply pick an answer that reminds them of what a teacher said in class, for example, keywords or vocabulary. Thus if popular misconceptions are not included as answer choices in test items, educators can easily be misled into believing that students have mastered a particular concept because students have chosen the correct answer. Such answer selections do not reflect the transformation that is the hallmark of conceptual change. In our opinion, multiple-choice items that do not employ distractors based on known misconceptions are of questionable utility for measuring conceptual understanding.

An example illuminating this issue is presented in Table 1, which shows two related items, one a traditional item (left) and the other specifically constructed as a DDMC item (right). The content concerns the reason for day and night, an elementary astronomical concept deemed as one of the "…most essential ideas which form the Earth conception (Nussbaum 1985, p. 90)." Day and night require an understanding of the Earth-Sun system, specifically both a qualitative and quantitative understanding of the Earth's spin and its orbit about the Sun.

The traditionally formulated item appeared on the 1969 National Assessment of Educational Progress (NAEP) and was administered to third graders (Schoon 1988). The DDMC item was developed for the Astronomy and Space Science Concept Inventory (ASSCI) field test administered to seventh and eighth grade students in the spring semester of their middle school Earth science course. One might imagine that older students would perform significantly better than third graders at knowing the reason for day and night. After all, they are completing a course in which typically one-quarter of the content is astronomy. More importantly, the older students have presumably learned the same concept previously, and one may assume that their conceptual thinking has grown more sophisticated, encompassing the most basic concepts as well as those more complex. However, third graders performed better on the NAEP day/night item than middle school students performed on the ASSCI day/night item.

Table 1. Similar items drawn from the 1969 National Assessment of Educational Progress and the ASSCI Field Test. Students can perform at artificially high levels on multiple-choice items if popular misconceptions are not included as choices. The correct choice is in bold for both items. Choice B in the ASSCI item is a strong misconception-based distractor, attracting nearly as many students (37%) as those who answered the item correctly (45%)

	NAEP, Grade 3		ASSCI Field Test, Grades 7 and 8	
	One reason that there is day and night on Earth is that the:	P-value	Scientists explain that we have night and day because:	P-value
A	**Earth turns.**	**0.81**	the Sun goes out.	0.04
B	Sun turns.	0.08	the Earth moves around the Sun.	0.37
C	Moon turns.	0.04	clouds block out the Sun's light.	0.03
D	Sun gets dark at night.	0.06	**the Earth turns on its axis.**	**0.45**
E	I don't know.	0.01	the Sun goes around the Earth.	0.10

The ASSCI item includes distractors based on misconceptions that had been documented repeatedly in clinical interviews with students. All other ASSCI items are constructed in the same manner. Similar misconceptions are illuminated in a comprehensive review of the relevant research literature. A study of second-grade students found that many knew that the Sun was "on the other side of the Earth" at night, but showed no clear preference for whether it was the Earth or the Sun that moved (Klein 1982). Both Baxter (1989) and Vosniadou and Brewer (1987) found that students have several naïve explanations for daily changes. The Sun is obscured (going behind a hill or covered by clouds), the Sun physically moves around the Earth, or the Earth orbits the Sun in 24 h (a conflation of the day and the year). Sadler (1992) investigated the popularity of these misconceptions in a study of 1414 students in grades 8–12, finding that 66% knew the reason for day and night. He established that the most popular misconception for day and night is that the Earth orbits the Sun in a day.

The NAEP item does not include the popular misconception involving the Earth's orbit in the cause for day and night. Hence, NAEP third graders perform better (81% correct) than ASSCI seventh and eighth graders (45% correct). The vast majority of the older students who answer the item incorrectly prefer answer "B" (i.e., the Earth moves around the Sun) (see Note 9), while none of the NAEP distractors appears to attract the majority of those who answer the item incorrectly. Any attempt to measure student understanding of the day/night concept would be seriously flawed in using an item similar to the NAEP item, in that the "Earth turns" choice encompasses both the scientifically correct answer and the dominant misconception. Using an ASSCI-like item would give a teacher valuable information about whether students have moved beyond this common misconception. We believe that studies of the effectiveness of curricula or pedagogies can be seriously compromised if they do not explicitly measure changes in student misconceptions using either open-ended items or distractor-driven multiple-choice items.

Failure to replace a misconception with a scientifically accurate concept can have negative consequences beyond understanding one concept or performance in one course. For example, a lack of understanding of the reason for day and night may well have serious repercussions for the comprehension of other ideas in astronomy. An understanding of the reason for day and night is a prerequisite for many concepts in introductory Earth science courses. Without a spatial model that properly distinguishes between orbiting and spinning (i.e., revolving and rotating, as many teachers describe it), other systems are nearly impossible to understand. Students without an accurate understanding of day and night generally do not gain an understanding of the motion of the Sun or the cause of the seasons (Sadler 1995). Moreover, astronomical discoveries on the forefront of science that can motivate students to study because of their timeliness (e.g., finding extrasolar planets by occultation or Doppler shift, measuring galactic rotation using spectroscopy, or the differential heating of planets) require understanding physical models that are similar to the basic concept of day and night.

3. METHODS

Much work has been done on the development of assessment instruments for college-level introductory astronomy. The Astronomy Diagnostic Test (ADT) incorporated some of our Project STAR items (see discussion below) into a larger sample of original items to create a multiple-choice instrument that reflects the content of introductory college level astronomy courses (Hufnagel 2001; Zeilik, Schau, and Mattern 1998). Other instruments have been developed that examine particular topics within astronomy: lunar phases (Dai and

Capie 1990; Lindell and Olsen 2002); light and spectroscopy (Bardar *et al.* 2005; Bardar *et al.* 2006); star properties (Bailey 2006); and the shape of the Earth and gravity (Sneider and Ohadi 1998). Our item inventory differs from these in that it deals with all relevant concepts at each grade level, as defined by the NRC and AAAS. Our instruments were developed to serve several purposes including the following:

- To establish the levels of student understanding and the prevalence of particular misconceptions both prior to and after instruction in relevant science courses.
- To measure conceptual change (using pre/post-test administration) in pre-college students as a result of instruction.
- To gauge teacher mastery of the concepts that they teach.
- To measure conceptual change in teachers as a result of gaining experience over time or as the result of professional development activities.
- To examine teachers' understanding of students' conceptions (by predicting item difficulty and common student incorrect responses).

By 2003, the Science Education Department at the Harvard-Smithsonian Center for Astrophysics (see Note 10) found itself uniquely positioned to develop an inventory of DDMC items and subsequent test forms linked to the NRC National Science Education Standards and AAAS Benchmarks for astronomy and space science for the NASA "Structure and Evolution of the Universe" Education Forum based within the Smithsonian Astrophysical Observatory.

The Science Education Department (SED) and its DDMC item development expertise both grew out of the development of the nation's first astronomy course developed with high school students as its primary target group. Project STAR (see Note 11) (Science Teaching through its Astronomical Roots), funded by NSF, began in 1985 by combining the knowledge and talents of astronomers and high school astronomy teachers to determine course content and methods and to create and test course materials. From its beginning, STAR was developed with students' astronomy misconceptions in mind. Our early work characterizing the astronomy misconceptions of precollege students resulted in the professional development video, *A Private Universe* (Schneps and Sadler 1988) (see Note 12). During STAR's six-year development, various types of evaluation were used to test the efficacy of its activities and of students' conceptual change related to the course's astronomy content. Evaluations were constructed after a thorough review of the published research literature on students' astronomy misconceptions. When STAR undertook nationwide field testing of the draft curriculum, a concise and easily administered method was needed to conduct pretesting and post-testing of students in STAR classrooms and control classrooms. The DDMC test was adopted as the best tool given the size of the sample and efficiency of scoring that many responses. A comprehensive 47-item test was subsequently developed and used (Sadler 1992). After STAR's completion, the DDMC test format was used to evaluate other SED curricula and professional development projects. Two members of the original STAR team remain with SED and have been joined by other specialists, including content experts and a psychometrician, to form the department's assessment team.

Building on the foundation of our Project STAR work, we embarked on the development of items for the grades K–12 astronomy content standards and benchmarks with support from NASA and the NSF. With the goal of producing an Astronomy and Space Science Concept Inventory (ASSCI) of tests for each grade band (K–4, 5–8, and 9–12) containing 25–30 items per test, we realized that a test-item bank many times that size had to be developed first. Only through extensive field testing could the psychometric properties of each item be established (primarily difficulty and discrimination). Well-performing items with a range of difficulty could then be combined and tested together in order to comprehensively measure the broad range of concepts at each grade level. Our development team followed eight steps to build our test item bank and develop the final instruments: (1) review and cataloging of relevant misconception literature; (2) standards interpretation and draft item construction; (3) expert review and validation; (4) pilot testing; (5) large scale validation; (6) item analysis; (7) construction of final test instruments; and (8) field testing of final test instruments. We describe each step in the following paragraphs.

3.1. Review and Catalog Relevant Misconception Literature

We began developing the item bank by identifying misconceptions related to astronomy through a thorough review of the misconception research literature covering the science concepts linked to all the standards. (The ASSCI covers four primary school, nine middle school, and seven high school astronomy and space science standards and benchmarks.) The literature review process included a search of both published findings and of unpublished research, such as graduate theses and dissertations.

3.2. Standards Interpretation and Draft Item Construction

The standards and benchmarks are typically condensed statements of knowledge, each the result of considerable discussion and deliberation by many specialists. To develop test items that can measure student understanding of a standard, the standard had to be interpreted in terms of specific knowledge for which test items can probe (see Note 13). Further, a student's understanding of a standard must be measured, in this method, by questions that can be posed in a multiple choice format and the answers to these questions must reveal the student's conceptual understanding. To facilitate item development, a 2 hour team meeting was held for participants to discuss a standard in terms of this type of demonstrated student knowledge, as well as relevant misconceptions from the literature that team members had read as a part of the first step. The discussion of a standard was always wide ranging as the team comprised content experts, experienced teachers, and assessment specialists. The result of such a meeting was a set of ideas and topics that broke the standard down into what we term its "component concepts." The team leader formulated this information into an outline, including references to documented misconceptions linked to the standard from step 1. Team members then used the outline for guidance to write draft items during the next week. The following week the team met again for 2 hours where members presented draft items and discussed how well they addressed the standard, incorporated misconceptions as distractors, and their initial readability and scientific accuracy. As a result of this vetting process, 20–40 draft items written by five individuals might be reduced to 10–15 items for a specific standard.

Starting with the K–4 standards and working through in grade sequence, we examined a K–4 standard and investigated how its topic was later addressed in the grades 5–8 and 9–12 standards. This grade sequential work model often helped clarify for us the intent of a given standard, especially for the K–4 standards, which set the stage for all future content focus.

3.3. Expert Review and Validation

After the draft items were vetted, surviving items were compiled into sets given to Harvard and Smithsonian scientists (known to only one team member) (see Note 14) for a review to ensure that there were no ambiguities or scientific inaccuracies. The scientists were also provided with a copy of the relevant standard so they were aware of the developers' objective. In addition to content related comments, scientists often suggested alternative wordings and some submitted their own draft items. In many cases, scientists objected to the inclusion of distractors that were drawn from the misconception literature, writing feedback such as "surely no one thinks this." It is interesting that even though many of these experts taught astronomy, they could still be unaware of the misconceptions of students. Once scientists had reviewed a set of items, the items and comments were discussed in a team meeting. It occurred often that some items had to be iterated a few times with scientists as we sought to create a useful test item that could be simultaneously understood by a fifth grader, for example, and be scientifically accurate, which often involved the use of qualifying language that complicated an item's reading level. Items that survived this iterative expert review were placed into a draft inventory, along with any sketches for graphs or other illustrations. These improved draft items were then sent to scientists outside the CfA (also known to only one team member) for external expert review. Scientists from outside of the CfA were used to ensure a fair sampling of scientists' views, as well as to validate our internal vetting and expert review processes. After iterating with these scientists in the same way as with CfA experts, draft items still in the inventory were sent to a reading specialist whose evaluation ensured that the reading level of each item was no higher than the target grade level, and preferably two or three grades lower. Keeping the readability of the item low increases the probability that any given item will act as a probe of scientific knowledge and not of reading ability. Items requiring graphs or artwork were sent to a technical illustrator, who also worked in an iterative fashion with the developers to ensure unambiguous figures. After an item's reading evaluation and artwork were done, it was put into the master item database for use in test form compilation. Typically, about half of the draft items that began step 3 were discarded during this step for a variety of reasons (see Note 15).

3.4. Pilot Testing

Once all standards and benchmarks for a grade band (e.g., K–4) had proceeded through steps 1–3, three pilot tests, one for each grade band, were constructed. All three tests contained a subset of items with the purpose of identifying six "anchor" or "core" items that could be included on all three tests in this subject. Pilot tests for grades 5–8 and 9–12 included such anchor items from earlier grade levels to provide a broad range of items for use in future tests. Anchor items allow for comparison of the overall performance levels of student groups on all test forms that contain them. Anchor items can also be used to standardize student performance if there

is significant variation among classrooms taking the tests. The pilot tests were constructed using 60 items selected from across the entire K–12 inventory—15 grades K–4 items, 36 grades 5–8 items, and 9 grades 9–12 items—to represent both the standards per grade band, as well as the most common misconceptions. Each pilot test contained 20 content items, the optimum length for student completion based on our earlier work designing similar tests for middle school physical science (see Note 16). Items were distributed proportionately across the forms to represent the standards from the three grade bands on each form, e.g., no test form had more middle school items than the others. Identical demographic items, such as gender, race, and age, were added to each form. Classrooms were recruited from around the country to secure upward of 100 teachers and 1000 students per pilot test, drawing upon a large cadre of teachers who volunteered for this project in response to a national mailing. Each pilot test was administered to students in grades 5–12 [the fifth graders served as proxies for K–4 students (see Note 17)]. Forms were mailed and administered by the teachers, and the returned answer sheets were logged and scanned.

Statistical and psychometric analysis of this core pilot study revealed estimates of difficulty, discrimination, and the popularity of the wrong answer(s) for each item. The variance explained by each pair of test items from a single pilot test (i.e., 400 combinations of 20 items) was calculated (Fig. 1). This variance is the fraction of total variation in total test score that is explained by only two items. Analysis of student performance on these items revealed the items' basic characteristics, as well as identifying poor items (e.g., apparent random selection of incorrect answers). The best pair of items typically explained greater than 60% of the variance on any one pilot test. The example in Fig. 1 shows that several pairs of items have a combined variance of greater than 0.60: items 2 and 7; 5 and 7; 5 and 20; 7 and 20. This analysis was carried out for each pilot test.

A total of six items from the three pilot tests—no two addressing the same standard (e.g., the solar system) to avoid overemphasizing any one standard—were selected for use as anchor items on all future test forms. The six anchor items were selected primarily on the basis of three criteria. These six items: (1) represented different standards, (2) accounted for more variance in the total test score than other items related to the standard, and (3) included one relatively powerful distractor as evidence of a misconception.

3.5. Large Scale Validation

Item	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	.44	.31	.45	.51	.36	.51	.31	.32	.44	.40	.31	.21	.34	.37	.36	.39	.40	.27	.49
2		.42	.53	.59	.46	.60	.49	.49	.55	.52	.45	.40	.46	.48	.49	.52	.54	.46	.59
3			.39	.50	.28	.51	.28	.33	.42	.38	.26	.16	.32	.29	.33	.35	.39	.25	.46
4				.58	.44	.58	.41	.44	.55	.50	.43	.33	.42	.42	.45	.53	.52	.42	.57
5					.50	.62	.49	.47	.57	.54	.50	.45	.52	.49	.51	.57	.56	.52	.63
6						.53	.32	.33	.44	.41	.25	.20	.33	.32	.36	.39	.38	.31	.50
7							.48	.49	.58	.58	.49	.43	.52	.49	.51	.57	.55	.50	.63
8								.31	.44	.38	.26	.17	.31	.28	.36	.39	.36	.27	.48
9									.44	.39	.28	.22	.37	.31	.36	.43	.42	.31	.48
10										.53	.42	.35	.47	.46	.48	.52	.51	.44	.56
11											.36	.35	.39	.41	.44	.51	.50	.38	.52
12												.14	.31	.27	.33	.37	.36	.25	.49
13													.24	.20	.27	.31	.28	.17	.46
14														.35	.38	.44	.40	.31	.50
15															.38	.44	.37	.33	.51
16																.42	.45	.32	.50
17																	.44	.38	.56
18																		.38	.56
19																			.47

Figure 1. Pilot test variance accounted for by two-item pairs. Each intersection of item numbers is the combined amount of variance in total pilot test score explained by each pair of items for ASSCI Pilot Test 3. Item pairs with variances greater than 0.60 (2 and 7; 5 and 7; 5 and 20; 7 and 20) are highlighted. Each pair is a good candidate for common "core" items to anchor all final test forms to be used for research into student and teacher conceptual understanding

Teachers were recruited from a national mailing to National Science Teachers Association members. Twenty-four elementary school teachers (with 1878 students), 33 middle school Earth and space science teachers (with 3763 students), and 31 high school Earth and space science or astronomy teachers (with 1958 students)

participated for a total of 88 teachers and 7599 students. Each field test instrument was measured for reliability (i.e., internal consistency) and validity (e.g., expert assessment of item scientific accuracy, match to NRC standards and AAAS benchmarks, and alignment with other test instruments).

Students in grades 7–12 were tested in April and May of the academic year 2003–2004. (The K–4 tests were given to fifth grade students during the fall of 2003, as near to the end of their K–4 experience as was feasible.) Therefore, it can be assumed that the results measure student understanding after nearly two semesters of study of the relevant grade band concepts. Teachers were also requested to select the correct answer for each test item as a way to determine their level of understanding of the concepts underlying the NRC standards. In addition, teachers were asked to predict the percentage of the students in their class who would answer each test item correctly, an attempt to measure the familiarity of teachers with their own students' knowledge level (Lightman and Sadler 1993). A summary of the ten tests appears in Table 2. They ranged in length from 25 to 27 items (with additional demographic questions at the end). Student means ranged 38–49% correct. Teacher means ranged 73–92% correct. KR-20, a measure of internal consistency (i.e., a measure of how well correct answers on individual items correlate with the total test score), was reasonable for student tests, but much lower for the relatively high performing teachers, as was expected. Measures of internal consistency reflect the degree of variance within a test form. Therefore, tests that yield high means and small standard deviations do not yield as high on reliability as do tests where there is more variability.

Table 2. Performance of teachers and students on the 10 field test forms. This table summarizes the ten field tests involving 7599 students of 88 teachers. The student mean scores on each test range from 0.38 to 0.48. The measure of internal consistency, KR20, is 0.64 or higher for each test of students. Teacher mean scores are much higher than those of students, as expected

			Students				Teachers			
Grade Band	**Test**	**Items**	**N**	**Mean**	**SD**	**KR20**	**N**	**Mean**	**SD**	**KR20**
K–4	101	25	1040	0.45	0.16	0.70	13	0.84	0.10	0.64
	102	26	838	0.48	0.16	0.69	11	0.74	0.10	0.55
5–8	201	27	508	0.44	0.17	0.76	6	0.73	0.13	0.80
	202	27	843	0.38	0.15	0.71	4	0.78	0.06	0.26
	203	26	682	0.41	0.16	0.72	6	0.84	0.13	0.79
	204	27	747	0.49	0.16	0.72	8	0.82	0.06	0.34
	205	25	983	0.38	0.14	0.64	9	0.88	0.15	0.87
9–12	301	26	595	0.48	0.19	0.80	10	0.88	0.14	0.84
	302	26	602	0.46	0.18	0.77	8	0.88	0.06	0.34
	303	26	761	0.47	0.16	0.73	13	0.92	0.06	0.41
All	Core	6	7599	0.59	0.24	0.44	88	0.96	0.11	0.63

3.6. Item Analysis

Characteristics of each item were calculated from large scale validation test data in order to select the best combination of item quality and coverage of all standards for use on each final test instrument. Test items are most often described by two parameters: difficulty (fraction correct) and discrimination (correlation of individual item scores with subjects' total test score). Figure 2 shows the distribution of these two parameters graphed by the three grade bands. When an item shows a positive and large discrimination, the students with the correct response on average scored higher on the total test score. A negative or zero-order discrimination means that more students with low total scores answered the item correctly than did their higher scoring peers. Items with the highest discrimination are typically answered correctly by 50–80% of the students. Many difficult items suffer rather low discrimination, but since the items cover all the standards, low discrimination items can be interpreted as representing concepts that are particularly difficult for students to master. Easy items also have low discrimination. The reason for this is purely a function of the math. When difficulty is 0.50, discrimination can be as high as 1.0, which means everyone who answered correctly scored above the total test score mean and everyone who answered incorrectly scored below that mean. Note that four grade 5–8 items have negative discrimination. Higher performing students chose the correct answer less often than lower performing students. Such items may either have structural problems (e.g., unclear wording) or may be measuring the popularity of certain misconceptions among higher performing students. Including items with negative

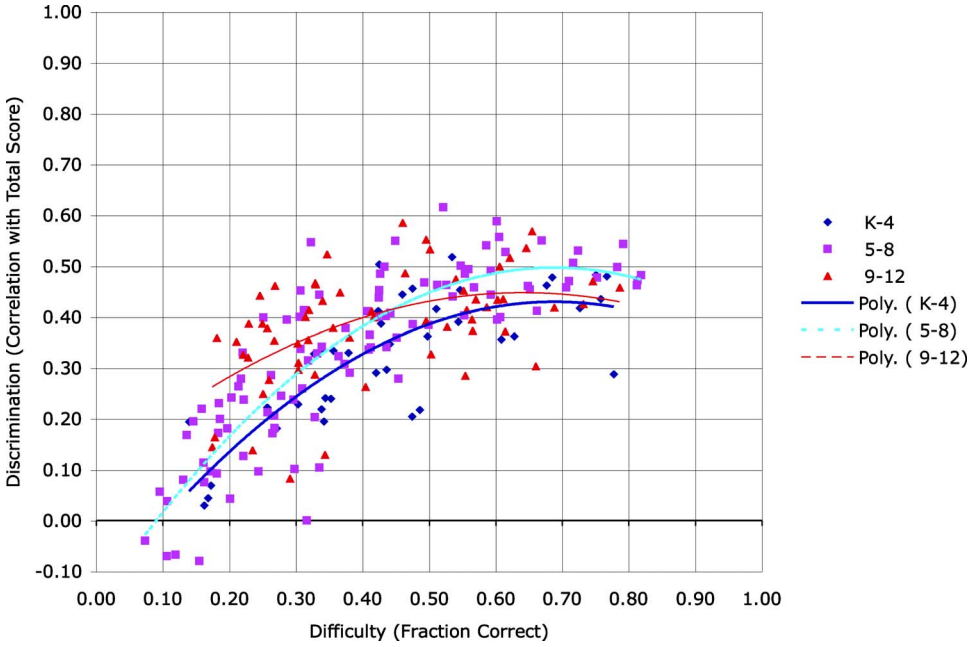

Figure 2. Difficulty and discrimination of test items by grade band. Results are similar by grade level, with maximum discrimination reached in the 0.50 to 0.80 difficulty level. Second order polynomials are fit to data at each grade level

discrimination in a test will lower its measures of internal consistency. However, such items may be very useful in a diagnostic test since they can identify misconceptions that are very popular among students who are learning a concept, but do not yet understand it completely.

Our team is particularly interested in a third statistic in addition to difficulty and discrimination, "misconception strength." DDMC items are constructed to gauge the appeal of particular distractors representing misconceptions derived from the research literature. The popularity of these ideas—often investigated using only qualitative methods or small-scale studies—has only rarely been measured in large, nationally representative groups of students (Sadler 1998). By analyzing the allure of each item's distractors, it is possible to gauge their relative popularity. Figure 3 shows the proportion of students choosing the most popular wrong answer out of the total number of students choosing any wrong answer as a function of item difficulty (see Note 18). Of the 211 items in the K–12 ASSCI test bank, 64 have a single distractor that attracted more than half of the students who answered the item incorrectly. In addition, the appeal of any particular misconception does not appear to be a function of item difficulty.

It is worth noting that even in high school, misconceptions are rife. As students age and mature, they continue to exhibit misconceptions about concepts they are studying. Holding a misconception prior to instruction

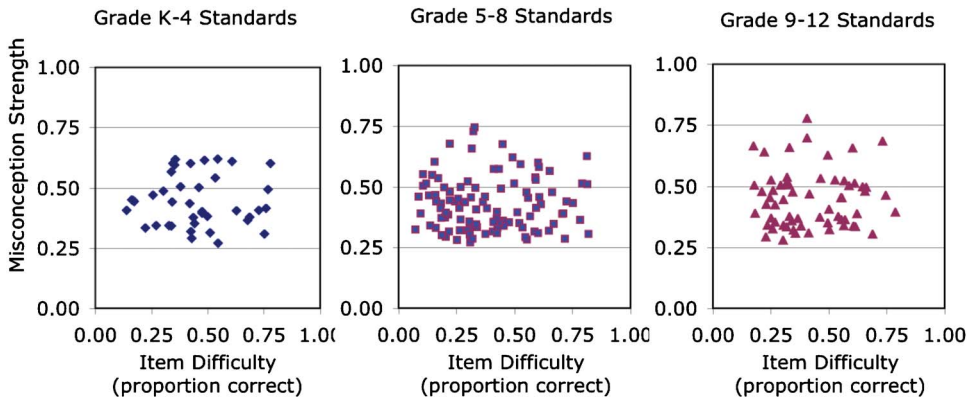

Figure 3. Distractor popularity versus item difficulty in each grade band. Test items for each grade band exhibit a range of item difficulty based on field tests. Appealing misconceptions exist for items at all levels of difficulty, exhibiting a misconception strength ≥0.50

appears to be a normal part of learning science at all grade levels. Thus, teachers at higher academic levels have the same problem as do teachers of younger students; misconceptions interfere with learning science.

3.7. Construction of Final Test Instruments

To conduct studies of conceptual understanding and change in target groups, the data from the field testing of 211 items were used to construct shorter test instruments that covered the standards for each of the three grade bands (see Note 19). Instrument development was done by first examining item difficulty, discrimination, and distractor popularity by standard, and then choosing items with the overall best profile for each standard within a grade band. For some standards, only a few items were in the final inventory due to removal of other items because we could not satisfactorily address issues raised during expert review of draft items. Items with the highest misconception strengths and greatest discrimination, distributed equitably over all standards and difficulty levels, were carefully chosen and collected to comprise the final instruments. By choosing items with a wide range of difficulty, we sought to have instruments that could accurately measure student understanding for a range of abilities and grade levels.

3.8. Field Testing of Final Test Instruments

The greatest need for these instruments was in assessing understanding of middle school and high school standards, so we concentrated additional effort in characterizing the final instruments for these two grade bands. The middle school test was administered to 787 students. The high school test was administered to 249 high school students and 145 college students taking introductory astronomy. Misconception strength and discrimination as a function of item difficulty for these groups are shown in Fig. 4.

We developed three academic level instruments, each having two forms (identical questions in differing orders). The elementary school instrument contained 12 items covering the K–4 standards. The middle school

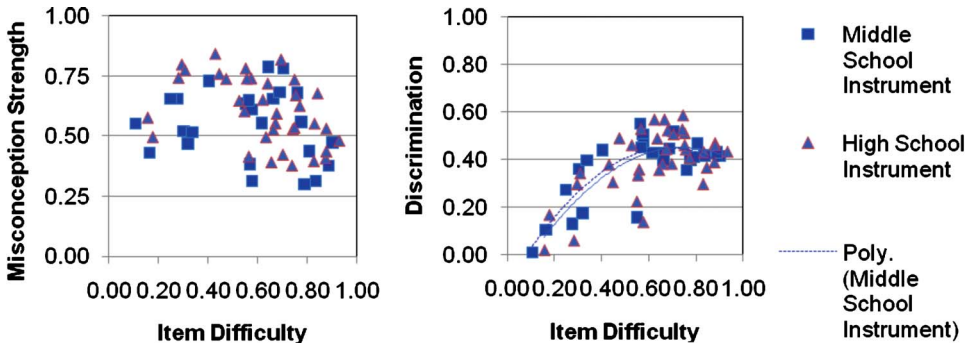

Figure 4. Distractor popularity (left) and discrimination (right) versus item difficulty for the final instruments. Note the high misconception strength of items for final instruments compared to all developed items (see Fig. 3). Items with negative discrimination were avoided

instrument contained 25 items: seven items from the grade K–4 standards to measure basic competency, 15 items from the grade 5–8 standards to broadly measure learning within the target grade band, and three items from the grade 9–12 standards to probe for students with advanced understanding. The high school instrument—which we also used to assess student conceptual knowledge in introductory college courses—included all the same items as the middle school instrument, as well as an additional 11 grade 9–12 items to focus the assessment on the high school standards and increase the difficulty of the test to avoid ceiling effects. Our resulting instruments displayed adequate reliability (KR-20) as measured by internal consistency when used with students in middle school Earth science, high school astronomy, and introductory college astronomy courses (Table 3). The difficulty of each test, as judged from post-test scores, was close to the ideal of the average overall test score being 50%, maximizing the overall discrimination of the test.

Table 3. Results from field-testing of final ASSCI instruments. Two instruments were assembled for final trials: A grade 5–8 instrument of 25 items and a high school instrument (also used with college students) consisting of the 25 grade 5–8 items plus 11 additional grade 9–12 items

Student Grade Level	N	Grade Band of Test Administered	Number of Items	Mean	SD	KR-20
6–8	787	5–8	25	0.51	0.16	0.73
9–12	249	9–12	36	0.61	0.18	0.85
College	145	9–12	36	0.65	0.15	0.80

4. ANALYSIS AND DISCUSSION

Comprehensive testing of 7599 students and their 88 teachers of our field testing allows for an estimation of the degree to which both students and teachers have mastered the standards in a nationally representative sample.

4.1. Conceptual Understanding of Teachers and Students

Table 4 presents the results for student and teacher performance on the standards within each of the three grade bands (K–4, 5–8, and 9–12). The number of items for each standard (last column) varies considerably because they include only those items that passed the vigorous scrutiny and final approval of our expert reviewers. Many initial items were eliminated from use because they were found to be unclear or dubiously accurate in their portrayal of the accepted scientific view, exhibited too high a reading level or had structural problems. We also found that writing items for some standards was very difficult and for others, easy, due to the science content of the standard, especially for some high school standards. Teachers both selected what they thought to be the correct answer for each item on their students' test and estimated the proportion of their own student who selected the same answer. The mean scores along with the standard error are presented in Table 4 for each of the standards examined.

Table 4. Performance and prediction on astronomy items for standards in each grade band. Descriptive phrases for the content of each standard are grouped by grade band. Student mean scores by standard ranged from a minimum of 0.29 correct to a high of 0.58. The minimum mean score for teachers was 0.67 and the maximum was 1.00. "Predictions" refers to how well teachers predicted their students' performance by averaged standard

Grade Band	Standard	Students		Predictions		Teachers		Number of Items
		Mean	SE	Mean	SE	Mean	SE	
K–4	There are different objects in the sky	0.39	0.00	0.51	0.03	0.74	0.03	26
	The sun heats the earth.	0.49	0.01	0.62	0.06	0.92	0.05	2
	There are many stars in the sky.	0.58	0.01	0.61	0.04	0.85	0.04	5
	Objects in the sky have patterns of motion	0.55	0.01	0.62	0.05	0.86	0.03	7
	Entire Grade Band	0.51	0.00	0.59	0.03	0.83	0.02	40
5–8	The solar system has a star, planets and other objects.	0.51	0.00	0.65	0.04	0.84	0.03	15
	Solar system objects move predictably	0.41	0.00	0.55	0.04	0.84	0.02	35
	Gravity is the key force in the solar system	0.40	0.01	0.47	0.04	0.67	0.07	10
	The sun's energy underlies many terrestrial phenomena	0.33	0.00	0.51	0.04	0.70	0.05	18
	Stars are fixed relative to each other	0.33	0.01	0.44	0.05	0.68	0.07	7
	Planets move relative to the stars	0.33	0.01	0.51	0.12	0.70	0.15	2
	Telescopes extend our vision.	0.52	0.01	0.64	0.04	0.89	0.04	9
	Stars are clustered in galaxies.	0.42	0.02	0.63	0.19	1.00	0.00	1
	Light takes time to travel.	0.36	0.01	0.53	0.04	0.87	0.04	7
	Entire Grade Band	0.41	0.00	0.54	0.03	0.79	0.02	104

Table 4. (*Continued.*)

Grade Band	Standard	Students		Predictions		Teachers		Number of Items
		Mean	SE	Mean	SE	Mean	SE	
9–12	The "big bang" theory	0.39	0.01	0.54	0.04	0.87	0.06	5
	Early star and galaxy formation	0.43	0.01	0.48	0.05	0.73	0.08	3
	Stellar fusion and its effects	0.29	0.01	0.55	0.04	0.88	0.04	9
	Stellar variation	0.54	0.01	0.60	0.04	0.94	0.02	10
	Light element formation	0.34	0.01	0.51	0.04	0.90	0.05	9
	Heavy element formation	0.48	0.01	0.57	0.04	0.84	0.05	5
	Obtaining and analyzing astrophysical data	0.44	0.01	0.54	0.03	0.91	0.02	26
	Entire Grade Band	0.42	0.00	0.54	0.03	0.87	0.02	67

Looking at student performance, it appears that we successfully used many of the popular misconceptions as distractors. Student mean score by standard fell into a range from 0.26 correct to 0.59 correct across the entire K–12 range. Teachers overpredicted their students' performance on every standard. For the grades K–4 standards, teachers overestimated student mean scores by 0.08. At the 5–8 grade level, teacher estimates were 0.13 too high. At the high school level, teachers predicted student performance, on average, to be 0.12 higher than actual. As a whole, teachers themselves did well on the test items, although for several standards, weaknesses are apparent. These data are presented in Fig. 5.

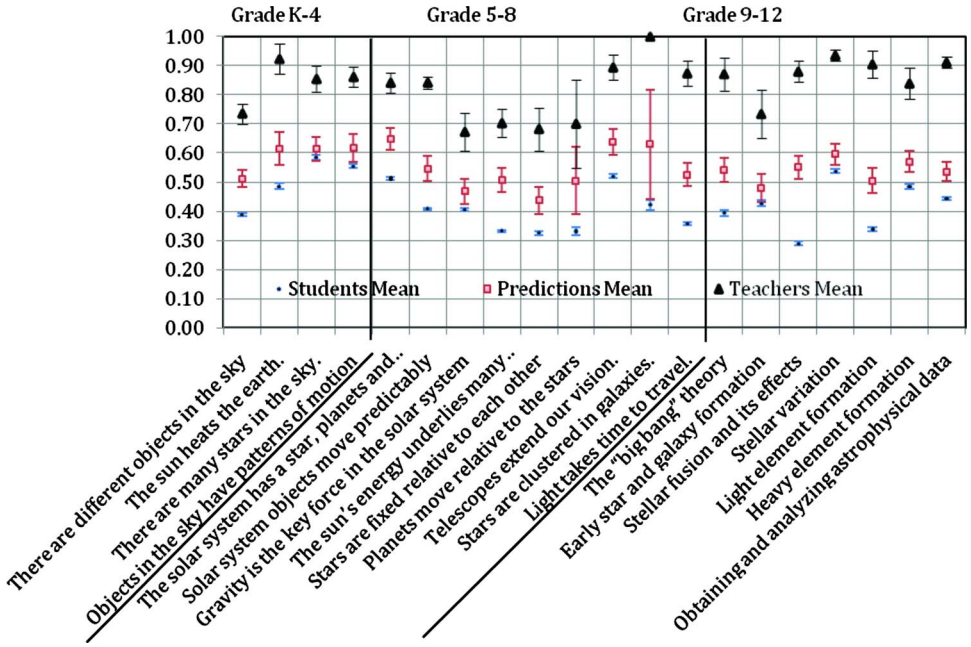

Figure 5. Student and teacher results for three grade bands. Overall, students do not show mastery (see Note 24) of any standard at their grade level. On average, teacher performance shows mastery, but with some gaps in knowledge. Teachers significantly overpredicted their students' performance, as seen by most teacher predictions being greater than 2 standard errors above student performance. Grade K–4 subjects included 1878 students of 24 teachers. Grade 5–8 subjects included 3769 students of 33 teachers. Grade 9–12 tests were administered to 1958 students of 31 teachers. Error bars show ±1SE

Differences in overall teacher performance between grade bands are not always significant, as seen by the overlapping of error bars representing the standard error of the mean. However, differences in student performance could be summarized as K–4 students performing significantly better on measures of grade-level understanding than students at higher grades in that their average level of mastery is higher than for students in later grades. Student and teacher performance indicate the level of understanding at the start of middle school (students at the start of grade 5 taking the test based on the K–4 standards); at the end of a middle school Earth science course in which astronomy is typically one-fourth of the content; and at the end of a high school Earth science or astronomy course. An examination of the most popular incorrect answers broken down by standard reveals those misconceptions that are still held by a large fraction of those students tested (Hewson and

Hewson 1983). We also found that some teachers hold some of the same misconceptions as their students. For each grade band, the relevant standards are listed below. We discuss all test results with a focus on the NRC standards because they comprise the majority of all K–12 astronomy and space science content guidance.

4.2. K–4 Grade Band

The NRC standards and AAAS benchmarks at this level are concerned only with celestial objects that are visible to the naked eye. By understanding their observable properties, positions and movements, students begin to build a knowledge base on which future course work will be based. The Sun has a special role in that it is the easiest object to observe and that it is the primary source of light and heat that affects the Earth's surface and its inhabitants.

4.2.1. NRC Standard: Objects in The Sky

The Sun, moon, stars, clouds, birds, and airplanes all have properties, locations, and movements that can be observed and described.

The Sun provides the light and heat necessary to maintain the temperature of the Earth.

4.2.2. NRC Standard: Changes in The Earth and Sky

Objects in the sky have patterns of movement. The Sun, for example, appears to move across the sky in the same way every day, but its path changes slowly over the seasons. The moon moves across the sky on a daily basis much like the Sun. The observable shape of the moon changes from day to day in a cycle that lasts about a month.

Listed below are grades K–4 major misconceptions persisting after completing fourth grade. Misconceptions also held by students' teachers are marked by an asterisk (*).

- The moon is always visible on clear nights and never in the daytime.
- The moon can be closer than the clouds.
- The period of the moon's phases is irregular or different than a month.
- The moon never has a gibbous shape (see Fig. 6).*
- There are no seasonal changes in the path of the Sun in the sky.
- The North Star does not hold a constant position (in North America).*
- All stars are white.
- The stars and moon do not move over the course of a night.*
- Volcanoes and human activity are seen as the primary heat sources for the atmosphere and oceans.

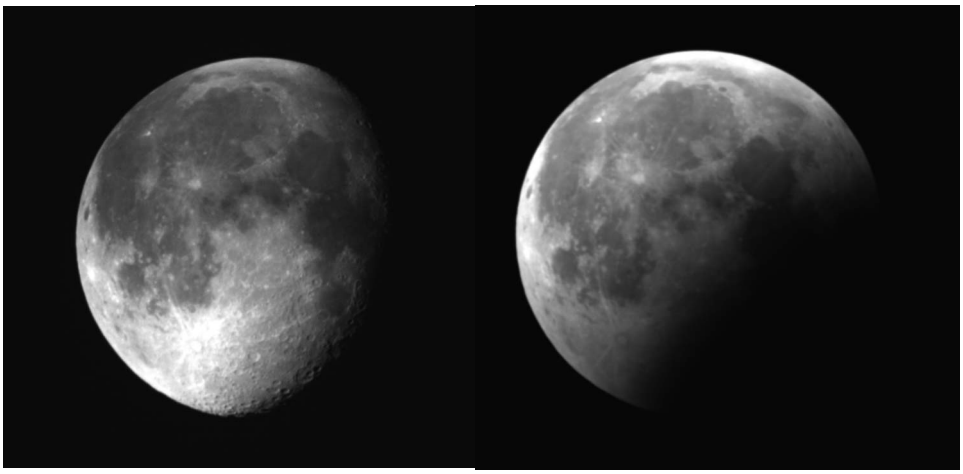

Figure 6. Comparison of shape between a gibbous phase (left) and a lunar eclipse (right). Both teachers and students prefer, in error, the image of the lunar eclipse over that of a gibbous moon, as representing a phase shape in the monthly change in the appearance of the moon. Courtesy MicroObservatory Telescope Project (http://mo-www.harvard.edu)

Since the emphasis of the NRC standards for grades K–4 is primarily observational, it is surprising to find that some of the misconceptions appear to be the result of a lack of observation. For example, a single careful scanning of the sky on a clear night will often reveal that the moon is nowhere to be seen. A classroom conversation concerning the visibility of the moon on the previous night can raise the issue of the moon not being visible. Young children may not have an opportunity to view the night sky over many evening hours, so the motion of the moon and stars, or the lack of motion of the North Star, may not be commonly observed. This situation is not an issue in the daytime when observations can be made easily during the school day. There is little impediment to observing the moon in the daytime sky during school hours. there are several times throughout the school day (lunch, recess, or even class time) when a group of students may be taken outside to make observations. Only one daytime sighting of the moon is necessary to bring into conflict students' views that the moon is *never* visible during the daytime. Tracing the shadow cast by a gnomon (a stick set vertically on a horizontal surface) outdoors or by a gnomon placed indoors by a sunny window is a common activity in several curricula (e.g., Project ARIES). Only a single day of gnomon shadow recording for any day of the year is required for students to observe that the Sun is not directly overhead at noon every day in the continental U.S. The fact that students possess misconceptions that can be challenged by such simple, direct, and personal observations suggests to us that few students make systematic, or even occasional, observations of the sky. Instead, we postulate that the study of astronomy at this grade level is primarily book-based memorizing of facts (e.g., memorization of the order and telescopic appearance of planets), with little learning by experience or evidence.

Teachers also appear to maintain several misconceptions, including an incorrect concept of the shapes of the lunar phases. Teachers do not seem to fully comprehend that the moon often assumes a gibbous shape and that the "bite out of the moon" shape characteristic of a lunar eclipse is quite rare (Fig. 6). This lack of understanding by teachers of the difference between a lunar eclipse that passes in a few hours and the daily change in phase during a month cannot help students acquire these concepts. Along with a lack of recognition of the position and constancy of the North Star and the nightly shift in other stars and the moon, teachers appear unfamiliar with night sky phenomena other than the presence of the stars and moon. Systematic observation of the pattern of motion of stars, the moon, the Sun, and planets would be a beneficial part of any professional development program aimed at increasing teachers' effectiveness in conveying astronomical concepts at this grade level. Keeping an observational journal of the sky over a period of time would be a useful and productive activity for teachers (Sadler, Haller, and Garfield 2000).

4.3. 5–8 Grade Band

The NRC standards at this level deal with the Earth-Sun-Moon system and related phenomena. The components of the solar system are defined and gravity is characterized as the primary force that controls its components' structure and movement. Student familiarity with the phenomena of day/night, seasons, moon phases, and eclipses is built upon presumed earlier observations in elementary school, as well as prior experience with models of motion and scale in the solar system. While many of the concepts to be learned at this level require an acceptance of astronomical facts for which students have little or no direct evidence, we believe that the construction and observation of very simple physical models that are accurate in scale, such as of the Earth-Sun-Moon system, may better help students understand many phenomena (e.g., moon phases and eclipses).

4.3.1. NRC Standard: Earth in the Solar System

The Earth is the third planet from the Sun in a system that includes the moon, the Sun, eight other planets and their moons, and smaller objects, such as asteroids and comets. The Sun, an average star, is the central and largest body in the solar system.

Most objects in the solar system are in regular and predictable motion. Those motions explain such phenomena as the day, the year, phases of the moon, and eclipses.

Gravity is the force that keeps planets in orbit around the Sun and governs the rest of the motion in the solar system. Gravity alone holds us to the Earth's surface and explains the phenomena of the tides.

The Sun is the major source of energy for phenomena on the Earth's surface, such as growth of plants, winds, ocean currents, and the water cycle. Seasons result from variations in the amount of the Sun's energy hitting the surface, due to the tilt of the Earth's rotation on its axis and the length of the day.

Below is a list of grades 5–8 major misconceptions persisting at the end of a middle school Earth science course. Misconceptions also held by students' teachers are marked by an asterisk (*).

- The Sun is not a star and not a member of the solar system.
- There are many stars within in the solar system.
- Stars other than the Sun are closer to us than Pluto.
- The Earth's orbit is highly elliptical. *
- The Earth turns on its axis once a year.
- The Earth orbits the Sun once a day, producing day and night.
- The space shuttle travels in the vicinity of the planets and/or stars.
- There is no gravity in space.

At this level, students exhibit many misconceptions concerning the Sun and the solar system. The Sun is seen as an exceptional, unique object, not a star like others in the universe. The scale and geometry of astronomical objects are misconstrued: stars are closer to the Earth than Pluto, the Earth's orbit around the Sun is highly elliptical, and the reach of humans into space is far greater than in reality. Orbits are not seen as a result of gravity because of a belief that there is no gravity in space. The causes of phenomena that are familiar to students—phases of the moon, day and night, and seasonal changes—are not understood. Instead, problems with scale are manifested with mental models in which the Earth, moon, and Sun are nearly the same size and only a few diameters of the Sun away from each other. Orbital and rotational periods are often not known or thought to be only an Earth day for all objects.

At this level, some teachers have a misconception about the shape of the Earth's orbit. They think that it is highly elliptical, although they do not think that this is the reason for seasons.

4.4. 9–12 Grade Band

At the high school level, students build on earlier concepts to extend their understanding of what scientists have discovered about the structure and history of the universe. Standards at this level require an acceptance of evidentiary data for systems that are vast in scale, well beyond the personal experience of students. Photographs, graphs, spectra, and other representations form the basis for acceptance of these remarkable discoveries and theories. Concepts from physics and chemistry provide a foundation for understanding star formation, the creation of elements, and the vast structures observed in the universe.

4.4.1. NRC Standard: The Origin and Evolution of the Universe

The origin of the universe remains one of the greatest questions in science. The "big bang" theory places the origin between 10 and 20 billion years ago, when the universe began in a hot dense state. According to this theory, the universe has been expanding ever since.

Early in the history of the universe, matter, primarily the light atoms hydrogen and helium, clumped together by gravitational attraction to form countless trillions of stars. Billions of galaxies, each of which is a gravitationally bound cluster of billions of stars, now form most of the visible mass in the universe.

Stars produce energy from nuclear reactions, primarily the fusion of hydrogen to form helium. These and other processes in stars have led to the formation of all the other elements.

Major misconceptions of grades 9–12 persisting at the end of a high school Earth science or astronomy course are as follows:

- The big bang created our solar system and all elements.
- Galaxies are held together by electromagnetism (not gravity).
- The universe is getting hotter.
- Probes have brought samples back to Earth from many planets.
- Astronauts have traveled beyond the moon.
- Red light is the most energetic wavelength.
- Telescopes are put in space to get closer to astronomical objects.

At the end of a high school course, many students think that the "big bang" formed the universe *as we see it now*. The current abundance of elements and the structures that we see today are thought to have existed from the very beginning of time. Gravity is not recognized as the primary force responsible for the large-scale

structures that are seen today, such as galaxies and star clusters. The reason for putting telescopes into space or on mountaintops is seen as simply decreasing the distance to astronomical objects; overcoming the effects of atmospheric distortion or absorption are deemed insignificant problems. Perhaps as a result of science fiction, many students think that astronauts have ventured far beyond the moon's orbit. Based on our findings of common misconceptions regarding the origin and evolution of the universe, the NASA Universe Education Forum initiated the Beyond the Solar System Professional Development project, which resulted in a DVD containing concrete, inquiry-based teaching resources for teachers to better address some of these concepts (see Note 20).

Few teachers held misconceptions at this grade level.

4.5. Teacher Performance and Predictions by Grade level

Although teachers perform far better than students on test items (refer back to Fig. 5), some test items prove far more difficult than others for teachers. We sought to determine if the items teachers found most difficult agreed with the items the students found to be most difficult. Figure 7 graphs student difficulty (fraction answering correctly, 1.00=100% correct) plotted against teacher difficulty by item. For each grade band, teachers perform well on the majority of items, with many items being answered correctly by all the teachers (teacher

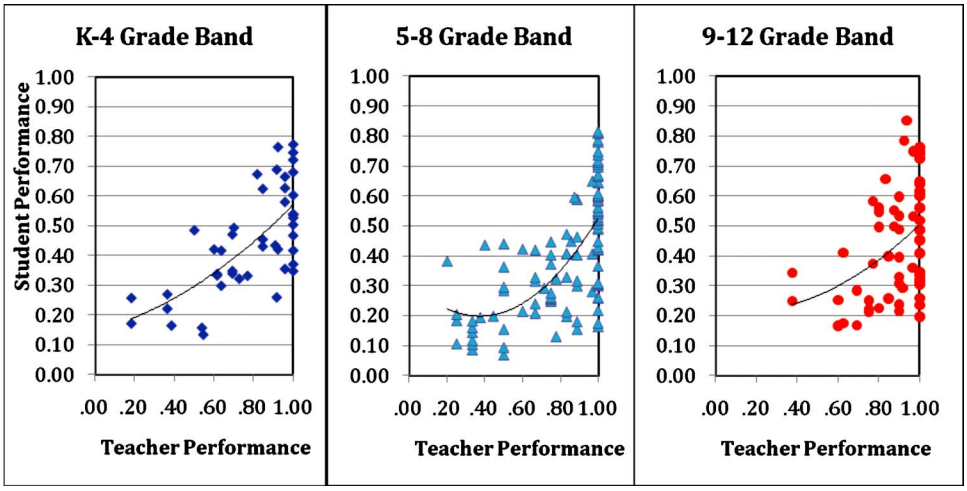

Figure 7. Comparison of item difficulty for students and teachers by grade band. Second order polynomials are fit to each graph. Student performance by item tracked with teacher performance: harder items for teachers were also harder for students. For items which were answered correctly by all teachers, only about half of students answered the question correctly, casting doubt on the belief that teacher subject matter knowledge is the sole explanation for poor student performance

performance=1.00). There are not many items for which teachers average less than 0.50 correct (5 at the primary school level, 13 at the middle school level, and 2 at the high school level). Perhaps the difficulty at the middle school level is the result of high school astronomy teachers acquiring stronger content background in their college studies. For the many items on which teachers performed perfectly (teacher performance=1.00), student performance averaged only about 0.50 correct (grades K–4: 0.56; grades 5–8: 0.54, grades 9–12: 0.48), thus bringing into question the notion that students' misconceptions stem from their teachers' lack of content knowledge or misconceptions. Even for items for which teachers have "perfect" content knowledge, student performance is not high at the end of a course. This calls into question whether efforts to increase teachers' subject matter knowledge through professional development will result in the desired increase in student learning. The fact that teachers can answer a test question correctly does not appear, by itself, to result in students learning that concept in their classroom (Leighton and Gierl 2007).

One possible explanation for low student performance on test items could be that teachers may overestimate either students' initial knowledge or students' ability to learn a concept, and thus not realize that they need to spend more time teaching a particular concept. Teacher overestimation can be the result of a teacher's lack of awareness of popular student misconceptions and how entrenched some of these misconceptions are in students' minds. This argument is buttressed by Fig. 8, which shows the relationship between teachers' predictions of student performance and students' actual performance by item. While data points appear widely

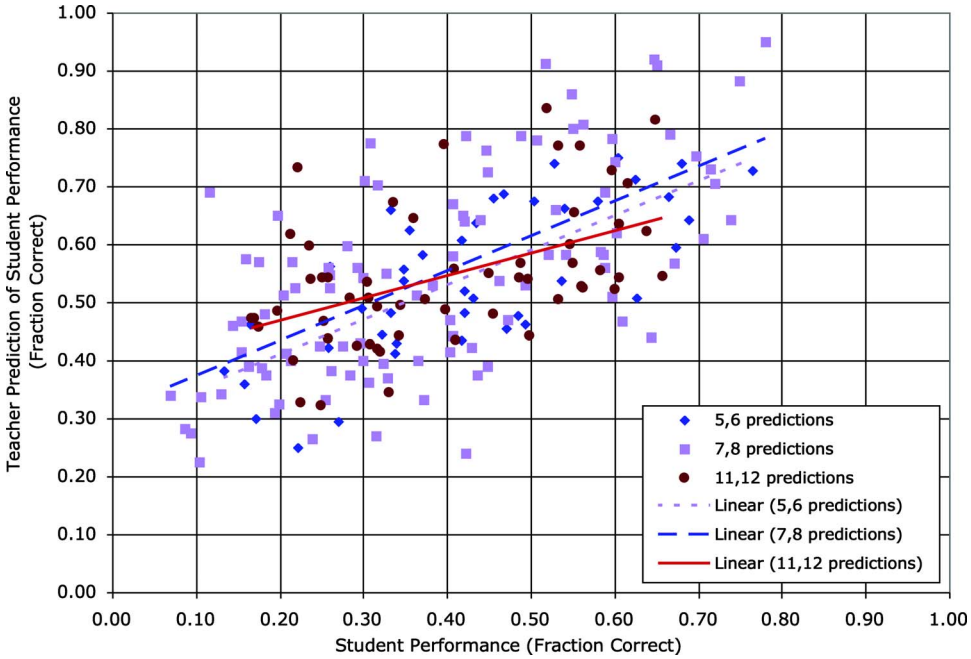

Figure 8. Relationship between student performance and teacher prediction of performance by grade band for test items. There is little difference between teachers' overestimation of the performance of their students for the three grade bands. This overestimation is particularly high for items for which students performed poorly. Teachers are more accurate in predicting the difficulty of items that students find easy

scattered, the best-fit lines, one for each grade level, mark the overall pattern, which is very similar for all three grade bands. Teachers generally overestimate the performance of their students as demonstrated by nearly all data points being above the solid black diagonal line that represents predictions matching student performance. On average, this overestimation is quite large. Teachers predict average scores of 0.55 for tests on which students only perform at the 0.40 level. For relatively easy items (student performance $>$0.50 correct), teacher predictions are much closer to student performance. For more difficult items (student performance $<$0.50 correct), the difference grows quite large. Teachers estimate a score of 0.45 on items which students answer correctly only 0.20 of the time (Wandersee 1986).

5. EXAMPLES OF THE USE OF ASSCI INSTRUMENTS FOR EVALUATION

The carefully constructed ASSCI instruments can be used as pretest and post-tests to measure change in performance as a result of a particular intervention or educational experience. We give two sets of examples. The first one shows the change in students' understanding of middle school and high school as a result of taking a course that covers astronomical content. The second example shows changes for four professional development institutes designed to improve teachers' subject matter knowledge in astronomy.

As a part of the NASA Universe Forum, we recruited a range of classrooms that had requested NASA materials and that taught either middle school earth and space science (at the seventh or eighth grade level) or a high school astronomy course (enrolling students in grades 9–12 as an elective). Seven hundred and eighty-seven middle school students of 15 teachers took a 25-item test that contained 7 K–4 items and 15 grade 5–8 items (plus 3 grade 9–12 items, which are not included in this analysis because of the small number of items for this higher grade band). Three hundred and ninety-seven high school students of 16 teachers had exactly the same items on their test, but also answered 14 grade 9–12 items. Both pretest and post-test included the same items, but in different orders. Each middle school teacher typically involved three to four classrooms in the testing, while high school teachers used one to two classes.

For middle school and high school students, our team was interested in finding whether students had mastered the standards of the prior grade band as well as the magnitude of gains for the particular concepts in that grade band. We hoped that students entering middle school had mastered the K–4 standards, and that students in high school astronomy classes has master both the K–4 and 5–8 standards. We assumed that the largest gains in knowledge would be at the middle school level for middle school students and at the high school level for high school students. Gains are calculated as "effect size" or ES, the difference between pretest and

post-test standardized in units of the pretest standard deviation. One value of using ES to examine gains is that it is a measure for which standards exist to gauge the magnitude of educational experiments (see Note 21).

Middle school earth and space science students scored at moderate levels (0.60) on the K–4 standards and showed no significant improvement on these standards during a year of study in which typically a quarter of the curriculum deals with astronomy (see Table 4 and Fig. 9). Middle school students gained a significant, but small amount on 5–8 standards. At the high school level, students began with only a moderate understanding of standards at the K–4 and 5–8 levels. They experienced small and significant gains on the K–4 standards, but no significant gains on the 5–8 standards. High school astronomy students achieved medium-sized gains on the 9–12 standards. The use of the ASSCI instruments for measuring student understanding revealed that these students did not undergo large changes in their understanding of astronomy concepts in school in the classrooms studied. High school astronomy courses appear to be most effective in increasing overall student understanding of concepts included in the NRC standards and AAAS benchmarks. It appears that teachers cannot rely on students having a firm foundation in concepts taught in earlier grades. It is unclear as to why the gains at the middle school level are small or why they are large at the high school level. Possible explanations are that the standards are too difficult for students, that high school astronomy courses concentrate more on astronomy than middle school classrooms, or other reasons. Opportunities exist to expand this research, using the ASSCI instruments, to help determine under what conditions high levels of gains are attained.

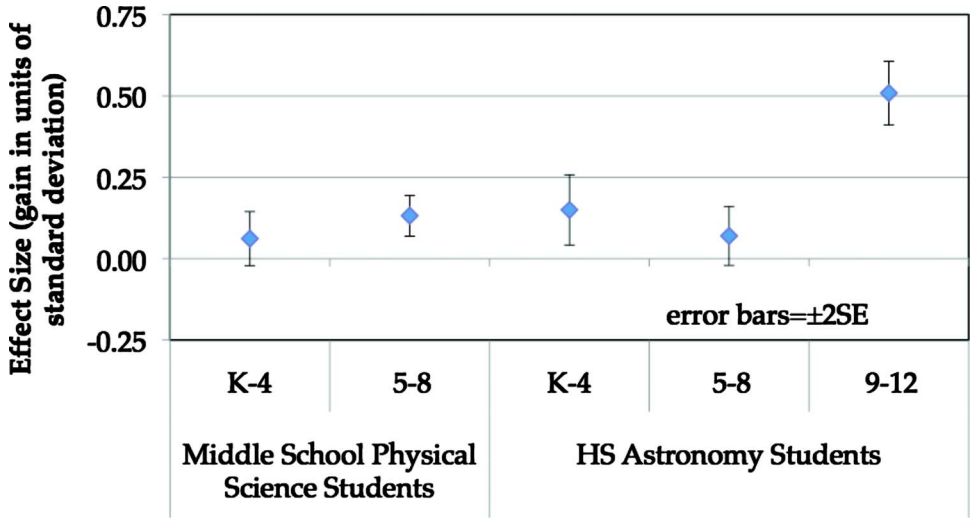

Figure 9. Effect size of middle school and high school classrooms. The only large gains seen were for high school students on grade band 9–12 standards. Bars showing ±2SE display whether or not the gains measured are significantly different from zero at the $p \leq 0.05$ level

All gains reported are small, except for high school astronomy students on grade band 9–12 standards. These students also experience a small gain on K–4 grade band standards for which they still score on only 67% of these items correctly by the end of their high school astronomy course.

Teachers are expected to increase in their understanding of science concepts through involvement in professional development activities. As a part of our MOSART grant (NSF-0412382), our team provides evaluation support to summer institute programs, four of which have focused on astronomy. These institutes ranged in length from 5 to 12 days and employed a range of options for increasing teachers' understanding of astronomy; learning to use particular curricula or activities, lectures from scientists involved in astronomical research, field trips to research facilities, working with scientists on current research, or conducting teacher-originated projects using research instruments (Table 5). While the number of institutes is too small to isolate the effects of these different approaches, we can calculate the gains for teachers by institute to see if there is a range of effectiveness or whether all institutes have roughly the same impact (Table 6). In all, 81 teachers attended these four institutes. Details are not presented for individual institutes to protect their anonymity. Each institute used instruments customized from the ASSCI item bank to match the particular concepts covered. Effect size can be effectively used to make comparisons across these different instruments.

Table 5. Results of pretest and post-testing middle school and high school students. Classrooms experienced gains for all grade bands, but significant gains were only found for grade 5–8 standards by middle school students and for grade bands K–4 and 9–12 for high school students. The last column in the table is a measure of the statistical significance of the gain

		Pretest			Post-test			Gain		
Course	Item Grade Band	Mean	SD	SE	Mean	SD	SE	ES	SE	t-test
MS Earth and Space Science	K–4	0.61	0.18	0.01	0.62	0.19	0.01	0.06	0.04	0.139
	5–8	0.53	0.18	0.01	0.56	0.19	0.01	0.13	0.03	0.000
HS Astronomy	K–4	0.64	0.18	0.01	0.67	0.19	0.01	0.15	0.05	0.006
	5–8	0.63	0.17	0.01	0.64	0.18	0.01	0.07	0.05	0.125
	9–12	0.48	0.2	0.01	0.59	0.22	0.01	0.51	0.05	0.000

Table 6. Astronomy professional development institute results. Institutes A and D had medium size gains. Institute C and D gains were not significant at the $P \leq 0.05$ level

	Pretest			Post-test			Gain		
Institute	mean	SD	SE	mean	SD	SE	ES	SE	t-test
A	0.65	0.16	0.03	0.74	0.16	0.03	0.57	0.11	0.000
B	0.74	0.16	0.04	0.77	0.14	0.04	0.15	0.16	0.888
C	0.83	0.17	0.05	0.85	0.12	0.03	0.12	0.14	0.395
D	0.87	0.10	0.02	0.91	0.08	0.02	0.47	0.15	0.002

Figure 10 graphically displays the range in effectiveness of the four institutes for which data was collected. Only two, institutes A and D, had significant gains showing that professional development can be either effective or ineffective at improving teacher subject matter knowledge. While this example is illustrative, other factors must be considered in carrying out a program evaluation. Typically, a regression model would be used to account for variance in pretest score, teacher background, and institute duration to control for differences between the teachers attending each institute.

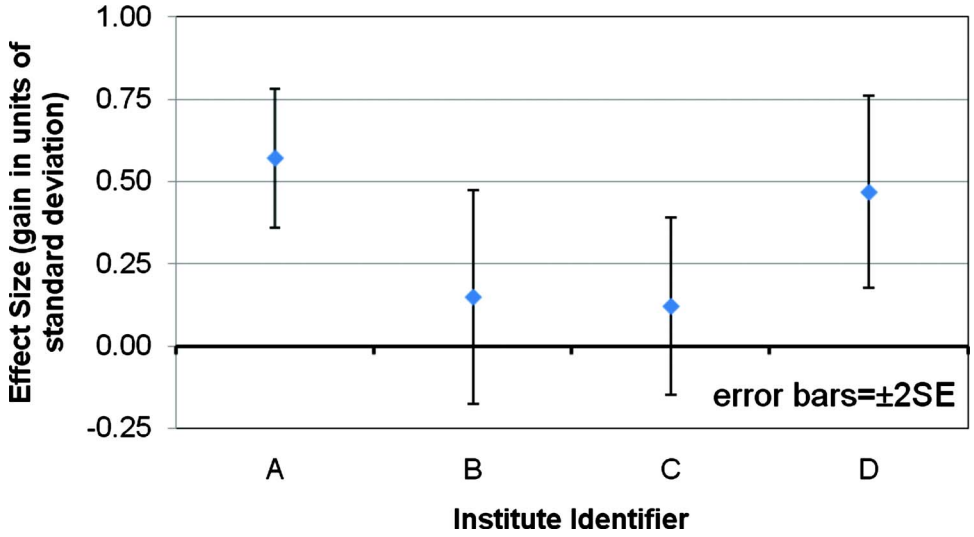

Figure 10. Effect Size of four astronomy professional development institutes. Only teachers at institute A and institute D were found to increase significantly on the second administration of the ASSCI instrument

Using instruments built from the ASSCI test bank to measure changes in subject matter knowledge (SMK) is one form of program evaluation. As we presented in Fig. 5, one can also ask teachers to predict the faction of their own students who can correctly answer an item. Teachers with knowledge of their students' capacities possess a kind of pedagogical content knowledge (PCK) that may make them more effective teachers (Shulman 1986). However, one must test each teachers' students in order to compare a teacher's predictions with his or her class performance. While we did not carry out such an extended study of these four institutes, we have considered that it may be possible to measure a key aspect of teachers' PCK without also testing their

students. We have found that for many items in our test bank, students across nearly all classrooms prefer one particular wrong answer out of the four distractors. These items, with misconception strengths greater than 0.50 (see Fig. 4), make up the majority of items on our final assessment instruments. Hence, the identification by a teacher of the most popular wrong answer—the dominant misconception—can be used as a measure of a type of PCK that is invariant across classrooms. While different fractions of students may answer an item correctly in different classrooms, the ones who answer it incorrectly appear to prefer a particular wrong answer. Teachers possessing the knowledge of which distractor students prefer would have the advantage of being able to teach in ways that would address particular student misconceptions. This would aid their students in reconstructing their understanding to better align with accepted scientific explanations. We have a research project underway to examine whether teachers who possess this kind of PCK, as well as SMK, experience greater gains in student understanding in science.

6. CONCLUSIONS

The standards developed by the National Research Council and the benchmarks generated by the American Association for the Advancement of Science represent the foundation for standards developed by each state. Under current Federal policy, each state is free to develop its own forms of assessment and testing schedule to monitor student progress. The development of these assessment systems is generally contracted to testing companies that rarely employ the research literature on student misconceptions as part of their item development. We have endeavored to develop a test bank that has items for each of the national standards and benchmarks at every grade level for K–12 astronomy and space science; each item was developed following an extensive review of the relevant research literature pertaining to misconceptions relating to the particular concept.

Our research has affirmed that many of the astronomy misconceptions found in the research literature are prevalent in precollege classrooms. Our data gathering has the benefit of characterizing the relative popularity of misconceptions that teachers may encounter in the different grade bands. This information is valuable since teachers generally overestimate the knowledge level of their students, may be unaware of student misconceptions and/or possess misconceptions themselves. We feel that this overestimation arises since teachers often write their own tests and quizzes, which generally do not engage students' misconceptions. Without the explicit inclusion of misconceptions, the ideas that students have constructed themselves to explain how the world works, teachers generally limit student responses to the available options in a multiple-choice item. Open-ended items are little better. While students may seemingly be free to respond drawing upon their misconceptions, they rarely do unless prompted to deal with common misconceptions. A telling example is an open-ended question asking for an explanation of seasons. Students will often draw the iconic diagram showing the tilt of the Earth as 23.5°, knowing that such specificity will be rewarded by the teacher. Instead, asking in which month the Earth is closest to the Sun will reveal whether they have a view that the seasons are caused by changing orbital distance. Teachers also have strengths and weaknesses, individually and as a group. When teachers have misconceptions, they are often the same ones that students espouse. Luckily, these instances are somewhat rare, but one should not assume that as teachers learned concepts during their schooling that they will remember the difficulties that they encountered. Teachers are surprisingly ill-informed as to the misconceptions of their students. It may be that teachers who do know their students' misconceptions can construct learning activities that are far more effective than those teachers who assume that their students are simply "blank slates" ready to absorb a particularly cogent elucidation of the scientific conception.

Developed by a team of educators and scientists, each item in the ASSCI has been validated by several content experts for clarity and accuracy. Item reading levels are appropriate to each grade band. A selection of items was initially pilot tested to select "anchor items" that appeared on all ten test forms of 25–27 items that were created for field testing. During field testing, a minimum of 500 students was used to collect data for each item tested. So that item parameters could be calculated, the middle school and high school tests were administered in April and May of the appropriate courses, after much of the astronomy content had been covered, while the K–4 field tests were scheduled for the beginning of the academic year for fifth graders to reflect learning up to grade 4. In all, 88 teachers took the tests themselves and predicted the performance of their own students (7599 total) on the test that they administered. Analysis shows that students found these DDMC tests rather difficult, with mean scores by grade band in the 0.40 to 0.50 level. At all levels, teachers performed relatively well, showing only a few consistent gaps in subject matter knowledge at the primary school, middle school, and high school levels. However, good teacher performance on items did not predict student mastery. Teachers generally dramatically overestimated their own students' performance, especially on items that were more difficult.

By selecting items from all the relevant standards and with a range of difficulty, final versions of the ASSCI were created and validated at the appropriate grade levels. The development and validation of short-form (12 items for primary school, 25 items for middle school, 35 items for high school) assessments covering this broad range of standards makes these instruments useful in the evaluation of astronomy curricula and teaching practices by testing students, extending the possibilities beyond studies that examine learning a single concept (Trundle, Atwood, and Christopher 2002; Trundle, Atwood, and Christopher 2007).

The assessments can also be used as a pretest to plan professional development offerings to address gaps in teacher knowledge since an emphasis on content knowledge is common in these efforts. Unlike studies that rely on teachers to subjectively report on the degree of increase in their own content knowledge to evaluate professional development programs (Garet *et al.* 2001), a pretest paired with a post-test (administered sometime after the conclusion of the professional development program) can objectively gauge the efficacy of teacher institutes and workshops. One particularly useful application of these tests is for teachers to administer them to their students after instruction, but prior to engaging in professional development. We have witnessed the tremendous impact on teachers finding that their own students still maintain certain misconceptions even after enthusiastic and engaging instruction. Such an experience can provide an opening for spirited discussion and motivate the study of more cognitively appropriate pedagogies and activities.

Teachers and professors can use these tests to determine the strengths and weaknesses of their students at the start of the term. While the teachers we surveyed had rather strong content knowledge, they lacked an accurate picture of the misconceptions held by their students. Studying the pretest performance of students can aid teachers in deciding on the appropriate activities to be used in their courses if they wish to positively impact conceptual understanding. An understanding of student misconceptions can also aid in determining which areas of students' conceptual foundations require strengthening, particularly if more conceptually sophisticated content will be covered in the course.

Public printable versions of these instruments are available at our self-service assessment website following the completion of a short tutorial on their use (see Note 22). Secure versions are available from the lead author for use in program evaluation. We hope to be able to offer secure, on-line administration of these tests in the future (see Note 23).

Generating the Astronomy and Space Science Concept Inventory required considerable effort on the part of staff and advisors, as well as the involvement of thousands of students and their teachers. On the way, its creators learned much about test development. No doubt, our own understanding of subject matter knowledge and pedagogical content knowledge has been strengthened. Our hope is that by creating assessment tools through this rigorous process others will not have to face the daunting task of creating such instruments. Instead, other researchers and educators can avail themselves of the opportunity to use these tools to improve their own teaching, to measure the effectiveness of different teaching methods and materials, and to evaluate the efficacy of professional development activities for those who teach astronomy and space science.

Acknowledgments

This work was carried out with support from the National Aeronautics and Space Administration's Structure and Evolution of the Universe Forum (Grant No. NCC5-706) and from the National Science Foundation grant for MOSART (Misconception Oriented Standards-based Assessment Resource for Teachers) (Grant No. NSF EHR-0412382). We wish to thank those who helped in item construction who are not listed as authors (Cynthia Crockett, Bruce Gregory, Jennifer Grier, and Bruce Ward) and the many astronomers who reviewed and commented on the items created. We appreciate the advice and support from Jeff Rosenthal and Larry Cooper of NASA, and of Irwin Shapiro and Charles Alcock of the Harvard-Smithsonian Center for Astrophysics. We greatly appreciate the involvement of teachers and their students in this project, without whom this research would have been impossible.

NOTES

Note 1: Charlesbridge Publishing, Watertown, MA.

Note 2: It's About Time, Armonk, NY.

Note 3: http://mo-www.harvard.edu/MicroObservatory/.

Note 4: For example, Vermont's statewide adoption of portfolio assessment was plagued by low inter-rater reliability (Koretz *et al.* 1994; Harlen 2005), as well as political controversy (Firestone 1998; Mathews 2004). Development of scoring rubrics and increased teacher training were augmented by use of outside authority to "standardize" teacher assessments and allow a fair comparison of scores. Even so, Vermont added a standardized test to increase reliability and lower the costs of assessing students (Vaishnav 2000). As of the date of this article, Vermont was using only standardized tests for school accountability.

Note 5: Modern attempts at identifying subjects' "alternative theories" or "scientific misconceptions" in astronomy and space science were first carried out in the domains of cosmography (Nussbaum and Novak 1976), light (Guesne 1978) and gravity (Gunstone and White 1981).

Note 6: Early efforts at developing multiple-choice tests around student misconceptions in astronomy included phases of the moon (Dai and Capie 1990), cosmology (Lightman and Miller 1989), cosmography (Nussbaum 1979), gravity (Ogar 1986), and compendia of concepts (Sadler 1987; Schoon 1988).

Note 7: Gorin (2006) proposed that diagnostic tests based upon of cognitive models would be very helpful to teachers in identifying the reasoning that students use. Morris *et al.* (2006) evaluated the psychometric properties of DDMC items and Ascalon *et al.* (2007) found that DDMC items are more difficult than open-ended items. Finally, Briggs *et al.* (2006) developed items for which each answer choice in is linked to a developmental level of student understanding, facilitating the diagnostic interpretation of student item responses. OMC items seek to provide greater diagnostic utility than typical multiple-choice items, while retaining their efficiency advantages.

Note 8: An example in astronomy concerns notions of the shape of the Earth. Nussbaum and Novak (1976) found that among second graders there is a popular notion that the Earth is ball shaped and that we live inside on a flat area with air held inside by a hemispherical shell. Repeated studies of American children have found that this misconception has remained prevalent over the 25 years since it was first discovered (Sneider and Pulos 1983; Sneider and Ohadi 1998). Similar views were later found across nations and cultures, including Nepal (Mali and Howe 1979); Israel (Nussbaum 1979); Greece (Vosniadou and Brewer 1987); among Han, New Zealand European, and New Zealand Maori children (Bryce and Blown 2006); and among Mexican-American children (Klein 1982).

Note 9: Scientists reviewing this item remarked that this particular distractor would have little appeal to students, instead predicting that answer E (the Sun goes around the Earth) would turn out to be the dominant misconception. Results showed otherwise.

Note 10: The CfA is a collaborative comprising Harvard College Observatory (HCO), Smithsonian Astrophysical Observatory (SAO) and the Astronomy Department of Harvard University. SED staff are affiliated with one or more CfA partners.

Note 11: Kendall/Hunt Publishing, Dubuque, IA.

Note 12: This video documents the misconceptions of students and the problems faced by teachers who wish to change them. The video won several major awards and is utilized extensively by teacher preparation and enhancement programs. With funding from the NSF and the Annenberg/CPB Foundation, this work was expanded to the *Minds of Our Own* series of three, 1 h documentaries aired on PBS during 1998 and associated teacher teleconferences. Demand from practitioners has led to a series of video-based professional development workshops (visit www.learner.org for more information).

Note 13: For example, a K–4 NRC Standard states that "(t)he sun provides the light and heat necessary to maintain the temperature of the Earth (National Research Council 1996, p. 134)."

Note 14: The identities of the reviewers were known to only one team member who maintained their anonymity so that they could feel free to comment candidly (which they often did!).

Note 15: It should be noted that once item development was underway, steps 2 and 3 occurred concurrently, with step 2 for a new standard occurring in parallel with step 3 for the prior standard. The development team's work on step 3 was interwoven as needed with step 2 actions. The work on steps 2–3 for the 20 K–12 astronomy and space science standards lasted 10 months.

Note 16: Although we first compiled a DDMC test for the evaluation of Project STAR, our work developing

DDMC items and tests based on the national standards began in 2001 under NSF grant REC-0087779, the Physical Science Assessment Project (PSAP). We chose to focus on the grades 5–8 standards because we had recently completed work developing a middle school physical science supplementary curriculum. The result of this work is the 110-item Middle School Physical Science Inventory. The work described in this article began upon the completion of PSAP.

Note 17: We used fifth graders as K–4 proxies because reading ability is more disparate among younger children, making reliable administration of written tests difficult. We are currently investigating alternative assessment methods for large-scale studies of younger children, such as through use of pictures or video with narration.

Note 18: Misconception strength equals the fraction of students choosing most popular wrong answer/(1–fraction choosing the correct answer).

Note 19: Public versions are available at: http://www.cfa.harvard.edu/smgphp/mosart/index.html.

Note 20: *Beyond the Solar System: Expanding the Universe in the Classroom—DVD*, Harvard-Smithsonian Center for Astrophysics, 2006. http://www.cfa.harvard.edu/seuforum/btss/.

Note 21: A value of 0.25 is considered small, 0.50 medium, and >0.75 large (Cohen 1988).

Note 22: http://www.cfa.harvard.edu/smgphp/mosart/ This site was developed and is maintained with funds from the NSF for Project MOSART.

Note 23: Availability of the online tests will be posted on the home page for Project MOSART II (NSF-0926272) on the NSF's Math and Science Partnership website at www.mspnet.org.

Note 24: We selected a score of 80% as the minimum performance level to represent mastery. We chose this score, equivalent to a B minus, based on the ubiquity across all academic levels in the U.S. with which such a grade is used to establish proficiency on state "high stakes" tests, a firm passing test grade and honors recognition.

References

Adams, J. P. and Slater, T. F. 2000, "Astronomy in the National Science Education Standards," *Journal of Geoscience Education,* 48, 39.

Arnaudin, M. W. and Mintzes, J. J. 1985, "Students' Alternative Conceptions of the Human Circulatory System: A Cross-Age Study," *Science Education,* 69, 721.

Ascalon, M. E., Meyers, L. S., Davis, B. W., and Smits, N. 2007, "Distractor Similarity and Item-Stem Structure: Effects on Item Difficulty," *Applied Measurement in Education,* 20, 153.

Ausubel, D. P., Novak, J. D., and Hanesian, H. 1978, *Educational Psychology: A Cognitive View*, New York: Holt, Rinehart, and Winston.

Bailey, J. M. 2006, "Development of a Concept Inventory to Assess Students' Understanding and Reasoning Difficulties About the Properties and Formation of Stars," Ph.D. dissertation, University of Arizona.

Bardar, E. M., Prather, E. E., Brecher, K., and Slater, T. F. 2005, "The Need for a Light and Spectroscopy Concept Inventory for Assessing Innovations in Introductory Astronomy Survey Courses," *Astronomy Education Review,* 4, 20.

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.

Baxter, J. 1989, "Children's Understanding of Familiar Astronomical Events," *Science Education,* 11, 502.

Briggs, D. C., Alonzo, A. C., Schwab, C., and Wilson, M. 2006, "Diagnostic Assessment with Ordered Multiple-Choice Items," *Educational Assessment,* 11, 33.

Bryce, T. G. K. and Blown, E. J. 2006, "Cultural Mediation of Children's Cosmologies: A Longitudinal Study

of the Astronomy Concepts of Chinese and New Zealand Children," *International Journal of Science Education,* 28, 1113.

Cohen, J. 1988, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Philadelphia: Lawrence Erlbaum Associates.

Dai, M. and Capie, W. 1990. "Misconceptions About the Moon Held by Preservice Teachers in Taiwan," 63rd Annual Meeting of the National Association for Research in Science Teaching (Atlanta).

Driver, R. and Easley, J. 1978, "Pupils and Paradigms: A Review of Literature Related to Concept Development in Adolescent Science Students," *Studies in Science Education,* 5, 61.

Duckworth, E. 1987, *The Having of Wonderful Ideas and Other Essays on Teaching and Learning*, New York: Teachers College Press.

Finley, F. N. 1986, "Evaluating Instructing: The Complementary Use of Clinical Interviews," *Journal of Research in Science Teaching,* 23, 635.

Firestone, W. A. 1998, "A tale of Two Tests: Tensions in Assessment Policy," *Assessment in Education: Principles, Policy & Practice,* 5, 175.

Freyberg, P. and Osborne, R. 1985, in *Learning in Science: The Implication of Children's Science*, ed. R. J. Osborne and P. Freyberg, Auckland, Heineman, 166.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., and Yoon, K. S. 2001, "What Makes Professional Development Effective? Results from a National Sample of Teachers," *American Educational Research Journal,* 38, 915.

Gilbert, J. K. 1977, "The Study of Student Misunderstandings in the Physical Sciences," *Research in Science Education,* 7, 165.

Gorin, J. S. 2006, "Test Design with Cognition in Mind," *Educational Measurement,* 25, 21.

Gould, R., Dussault, M., and Sadler, P. 2007, "What's Educational About Online Telescopes? Evaluating Ten Years of MicroObservatory," *Astronomy Education Review,* 5, 127.

Guesne, E. 1978, *Physics Teaching in Schools*, ed. G. Delacote, London: Taylor and Francis, 265.

Gunstone, R. F. and White, R. T. 1981, "Understanding of Gravity," *Science Education,* 65, 291.

Halloun, I. A. and Hestenes, D. 1985, "The Initial Knowledge State of College Physics Students," *American Journal of Physics,* 53, 1043.

Harlen, W. 2005, "Trusting Teachers' Judgment: Research Evidence of the Reliability and Validity Of Teachers' Assessment Used for Summative Purposes," *Research Papers in Education,* 20, 245.

Hewson, M. and Hewson, P. W. 1983, "Effect of Instruction Using Students' Prior Knowledge and Conceptual Change Strategies on Science Learning," *Journal of Research in Science Teaching,* 20, 731.

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

Kamen, M. 1996, "A Teacher's Implementation of Authentic Assessment in an Elementary Science Classroom," *Journal of Research in Science Teaching,* 33, 859.

Kaufman, J. 1997, "NASA in Crisis: The Space Agency's Public Relations Efforts Regarding the Hubble Space Telescope," *Public Relations Review,* 23, 1.

Klein, C. A. 1982, "Children's Concepts of the Earth and the Sun: A Cross Cultural Study," *Science Education,* 66, 95.

Koretz, D., Stecher, B., Klein, S., and McCaffrey, D. 1994, "The Vermont Portfolio Assessment Program: Findings and Implications," *Education Measurement: Issues and Practice,* 13(3), 133.

Leighton, J. P. and Gierl, M. J. 2007, *Cognitive Diagnostic Assessment*, Cambridge, Cambridge University Press.

Lightman, A. and Sadler, P. M. 1993, "Teacher Predictions Versus Actual Student Gains," *The Physics Teacher,* 31, 162.

Lightman, A. P. and Miller, J. D. 1989, "Contemporary Cosmological Beliefs," *JSTOR: Social Studies of Science,* 19, 127.

Lindell, R. and Olsen, J. 2002. "Developing the Lunar Phases Concept Inventory," in *Proceedings of the 2002 Physics Education Research Conference*, eds. S. Franklin, J. Marx, and K. Cummings, New York: PERC Publishing.

Mali, G. and Howe, A. 1979, "Development of Earth and Gravity Concepts Among Nepali Children," *Science Education,* 64, 213.

Mathews, J. 2004, "Portfolio Assessment: Can It be Used to Hold Schools Accountable?" *Education Next,* 4, 72.

Mintzes, J., Wandersee, J., and Novak, J. 2005, *Assessing Science Understanding*, Oxford: Elsevier.

Morris, G. A., Branum-Martin, L., Harshman, N., Baker, S. D., Mazur, E., Dutta, S., Mzoughi, T., and McCauley, V. 2006, "Testing the Test: Item Response Curves and Test Quality," *American Journal of Physics,* 74, 449.

National Research Council 1996, *National Science Education Standards*, Washington: National Academy Press.

Novak, J. D. 1998. *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations*. Mahwah: Lawrence Erlbaum Associates.

Nunnally, J. C. 1964, *Educational Measurement and Evaluation*, New York: McGraw-Hill.

Nussbaum, J. 1979, "Children's conception of the Earth as a Cosmic Body: A Cross Age Study," *Science Education,* 63, 83.

Nussbaum, J. 1985, in *Children's Ideas in Science*, eds. R. Driver, E. Guesne, and A. Tiberhien, Philadelphia: Open University Press, 170.

Nussbaum, J. and Novak, J. 1976, "An Assessment of Children's Concepts of the Earth Utilizing Structured Interviews," *Science Education,* 60, 535.

Ogar, J. 1986, "Ideas About Physical Phenomena in Spaceships Among Students and Pupils," in *GIREP—Cosmos and Educational Challenge*, Copenhagen: European Space Agency, 375.

Piaget, J. and Inhelder, B. 1967, *The Child's Conception of Space*, New York: W.W. Norton.

Prather, J. P. 1985. *Philosophical Examination of the Problem of Unlearning of Incorrect Science Concepts*, National Association for Research in Science Teaching ERIC document: ED256570.

Project 2061, 2001, *Atlas of Scientific Literacy*, Washington: American Association for the Advancement of Science.

Sadler, P. M. 1987, in *Second International Seminar on Misconception and Educational Strategies in Science and Mathematics in Ithaca, NY*, ed. J. D. Novak, Ithaca: Cornell University Press, 422.

Sadler, P. M. 1992, "The Initial Knowledge State of High School Astronomy Students," Ed.D. dissertation, Harvard University.

Sadler, P. M. 1995, in *Astronomy Education: Current Developments, Future Coordination*, ed. J. Percy, San Francisco: Astronomical Society of the Pacific, 46.

Sadler, P. M. 1998, "Psychometric Models of Student Conceptions in Science: Reconciling Qualitative Studies and Distractor-Driven Assessment Instruments," *Journal of Research in Science Teaching,* 35, 265.

Sadler, P. M., Haller, D., and Garfield, E. 2000, "Observational Journals: An Aid to Sky Watching," *Journal of College Science Teaching,* 29, 245.

Schneps, M. H. and Sadler, P. M. 1988, *Video: A Private Universe*, Santa Monica: Pyramid Films.

Schoon, K. J. 1988, "Misconceptions in Earth and Space Sciences: A Cross-Age Study," Ph.D. dissertation, Loyola University.

Shulman, L. 1986, "Those Who Understand: Knowledge Growth in Teaching," *Educational Researcher,* 15, 4.

Slater, T. F. 1997, "The Effectiveness of Portfolio Assessments in Science," *Journal of College Science Teaching,* 26, 315.

Sneider, C. and Pulos, S. 1983, "Children's Cosmographies: Understanding the Earth's Shape and Gravity," *Science Education,* 67, 205.

Sneider, C. I. and Ohadi, M. M. 1998, "Unraveling Students' Misconceptions about the Earth's Shape And Gravity," *Science Education,* 82, 265.

Trundle, K. C., Atwood, K. R., and Christopher, J. E. 2002, "Preservice Elementary Teachers' Conceptions of Moon Phases Before and After Instruction," *Journal of Research in Science Teaching,* 39, 633.

Trundle, K. C., Atwood, K. R., and Christopher, J. E. 2007, "Fourth-Grade Elementary Students' Conceptions of Standards-Based Lunar Concepts," *International Journal of Science Education,* 29, 595.

Turkle, S. 2008, *Falling for Science*, Cambridge: MIT Press.

Vaishnav, A. 2000, "Portfolios seen as partner to MCAS," *The Boston Globe, 24 May: B1, B5.*

Vosniadou, S. and Brewer, W. F. 1987, "Theories of Knowledge Restructuring in Development," *Review of Educational Research,* 57, 51.

Wandersee, J. H. 1986, "Can the History of Science Help Science Educators Anticipate Students' Misconceptions?" *Journal of Research in Science Teaching,* 23, 581.

Ward, R. B., Sadler, P. M., and Shapiro, I. I. 2007, "Learning Science Through Astronomy Activities: A Comparison Between Constructivist and Traditional Approaches in Grades 3–6," *Astronomy Education Review,* 6, 1.

Zeilik, M., Schau, C., and Mattern, N. 1998, "Misconceptions and Their Change in University-Level Courses," *The Physics Teacher,* 36, 104.

ÆR
010111-1–010111-26