JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 35, NO. 3, PP. 265–296 (1998)

Psychometric Models of Student Conceptions in Science: Reconciling Qualitative Studies and Distractor-Driven Assessment Instruments

Philip M. Sadler

Harvard University Graduate School of Education and Harvard Smithsonian Center for Astrophysics, Longfellow 315, Cambridge, Massachusetts 02138

Received 30 May 1996; revised 25 September 1997; accepted 28 October 1997

Abstract: We stand poised to marry the fruits of qualitative research on children's conceptions with the machinery of psychometrics. This merger allows us to build upon studies of limited groups of subjects to generalize to the larger population of learners. This is accomplished by reformulating multiple choice tests to reflect gains in understanding cognitive development. This study uses psychometric modeling to rank the appeal of a variety of children's astronomical ideas on a single uniform scale. Alternative conceptions are captured in test items with highly attractive multiple choice distractors administered twice to 1250 8th through 12th-grade students at the start and end of their introductory astronomy courses. For such items, an unusual psychometric profile is observed—instruction appears to strengthen support for alternative conceptions before this preference eventually declines. This lends support to the view that such ideas may actually be markers of progress toward scientific understanding and are not impediments to learning. This method of analysis reveals the ages at which certain conceptions are more prevalent than others, aiding developers and practitioners in matching curriculum to student grade levels. This kind of instrument, in which distractors match common student ideas, has a profoundly different psychometric profile from conventional tests and exposes the weakness evident in conventional standardized tests. Distractor-driven multiple choice tests combine the richness of qualitative research with the power of quantitative assessment, measuring conceptual change along a single uniform dimension. © 1998 John Wiley & Sons, Inc. J Res Sci Teach 35: 265–296, 1998.

The past decade of research in science education has been distinguished by an explosion of research into children's ideas. These efforts have revealed that most students hold conceptions that are quite different from their teachers' or those of scientists and that these ideas are resistant to change, even in the face of the most dogged efforts. These "misconceptions" or "alternative conceptions" have been uncovered using several research techniques; but qualitative interviews, traceable to Piaget, have proven incredibly productive (Duckworth, 1987; Osborne & Gilbert, 1980). Hundreds of papers have been published revealing student conceptions in a variety of domains (Pfunt & Duit, 1994).

Contract grant sponsor: NSF
Contract grant numbers: MDR 85-50297, MDR 88-50424
Contract grant sponsor: Smithsonian Institution
Contract grant sponsor: Apple Computer

 CCC 0022-4308/98/030265-32

This is not lost on teachers of science, who struggle to make use of the variety and volume of student ideas that they encounter. Is there value in getting everyone to formulate and express their own views? Should alternative notions be validated in class discussion? Do all students who come to understand scientific concepts previously hold certain alternative conceptions? What anchoring concepts can be built upon? Is there a more productive or optimal ordering of topics?

Many of us who teach science have trouble reconciling student ideas with the daily enterprise of teaching. Often, we cannot help but view students' ideas as oddities. Strangely familiar, these notions invoke smiles as often as knitted brows, like the youngster who after reading about the planets asked to see the earth through my telescope (Vosniadou & Brewer, 1987). Reading about children's views of the earth as hollow (Nussbaum, 1979) was not adequate preparation for my own second grader's explanation that spaceships must blast a hole in the sky to get into space, because we live inside the globe; I was dumbfounded. As teachers of science, we have personally long abandoned such notions and find it nearly impossible to identify with them or truly comprehend the grip they have on our students' minds. It is far easier to explain them away as the effect of misinformed teachers at lower grade levels, parents, or television than to integrate them as an essential and central step in the learning of science.

How can these ideas best serve us as teachers? The task is not simply to identify more alternative conceptions but to begin to relate them to what is familiar, the scientific ideas that have had a historically painful birth and which we now hope students will make their own. Several approaches have resulted in progress toward this goal, but all are underpinned by a notion of progression, that there are better and worse ways of arranging the experiences which we place before students (Millar, Gott, Lubben, & Duggan, 1993). Many researchers advise that students' alternative conceptions be studied and dealt with explicitly in teaching scientific concepts (Bell, Brekke, & Swan, 1987; Eaton, 1984; Hestenes, Wells, & Swackhammer, 1992; Nussbaum & Novak, 1982; Smith, DiSessa, & Rochelle, 1993). Several have attempted to place individual misconceptions within larger contexts. These "alternative frameworks" are used to generate experiences that challenge students' ideas (Driver, 1973). Others see alternative conceptions as a symptom of other abstract cognitive structures, termed "phenomenological primitives" or "p-prims" (diSessa, 1993; Hammer, 1995). Questions or scenarios cause students to draw upon primitives, such as "closer means stronger" explaining the cause of the seasons. Another popular view is that children's learning recapitulates the same progression as the historical development of the field and that repeating this sequence is beneficial for students (Wandersee, 1986). This is very appealing since there are great similarities, for example, in students' views of mechanics and those of Aristotle (diSessa, 1982). Moreover, the cognitive shifts from one conceptual framework to another can be compared to the "paradigm shifts" in the history of science (Kuhn, 1970). Such shifts can be caused by "discrepant events" that are in conflict with the current paradigm (e.g., Becquerel's discovery of radiation in 1896: uranium salts fogged a photographic plate in total darkness) or an inability to frame an outstanding problem in terms of the current paradigm (e.g., Einstein's quantization of light in 1905: explaining the photoelectric effect, which Maxwell's equations could not model). Each model—alternative framework, explicit, p-prim, optimal sequencing, or historical—suggests pedagogical or curricular innovations.

The greater task is to construct a sequential model for how students come to understand such scientific concepts. For this purpose, special attention must be paid to assessments that might capture the mental models of students. Standardized tests are often criticized for testing facts but not concepts, just bits of ideas but not their entirety. Yet, such tests offer hope in connecting the ways that novices think with the ways that experts think (Dufresne, Gerace, Hardi-

man, & Mestre, 1986; Hardiman, Dufresne, & Gerace, 1986; Reif, 1987). Such tests can uncover not only how near students' views are to those of experts, but how closely they cluster to a few alternative paradigms. Psychometrics, the measurement of mental processes through testing, offers an array of tantalizing tools to study conceptual change. Large-scale quantitative testing could help to uncover patterns, generalizing findings despite the uncertainty inherent in research on learning (Mislevy, 1994). The purposes of this article were to explore the relationship between scientific conceptions and alternative conceptions in the domain of astronomy and to use a method for mapping these concepts to a single progressive knowledge scale.

Background

Assessments of Alternative Conceptions

Interviews have revealed most of what we know about students' ideas in science. From astronomy to zoology, qualitative studies have revealed youngsters' understandings in a way that is both illuminating and believable. Moreover, the power of a student passionately explaining her ideas is far more appealing and convincing than dry, quantitative data, as I have found by presenting workshops for teachers at conferences and by participating in a national teleconference on student ideas.[1] The videotaped interviews of a few students who remind us of our own tend to be much more persuasive than the written tests of thousands. As a result, my workshops now start with the video *A Private Universe* (Schneps & Sadler, 1988).

Every methodology has its strengths and shortcomings. The major weakness of investigating student alternative conceptions through interviews is that these investigations, because of their one-to-one nature, rely on relatively few subjects. Large studies with greater than 100 subjects are rare in the literature. Because of their small size, these investigations are difficult to generalize to larger populations. Also, interviews must be conducted by those who are adept. Even so, interviewer bias may taint the findings. Yet, while qualitative methods have been and continue to be the most productive way of investigating children's ideas in science, multiple choice tests offer some benefits. Although written tests do not allow the interviewer to pursue the subject's ideas until they are clearly captured and ambiguous responses clarified, they ensure that all subjects are examined using identical questions. Moreover, paper and pencil tests are cheap and easy to administer and score.

Moving to the classroom, experienced teachers display an uncanny ability to predict students' alternative conceptions and to characterize the knowledge state of their students (Bjar, 1993; Borko & Livingston, 1989). Yet, this ability is limited to those with both mastery of the subject matter that they teach and the faculty for being able to listen carefully to their students. For novices, those teaching in a new field, or teachers who are unaware of their students' ideas, an easy-to-administer method is desirable. Ideally, simple diagnostic techniques to reveal student ideas would be of great value for practicing teachers. Haladyna (1994) pointed out that multiple choice tests which "pinpoint" alternative conceptions are useful to both teachers and students alike. In the same vein, psychometricians have begun to consider how to probe the cognitive processes that underlie test responses (Mislevy & Verhelst, 1990).

Written Tests of Student Conceptions

Particularly promising are written tests which offer a choice between a single correct answer and several alternative conceptions that have been identified through student interviews

(Freyberg & Osborne, 1985). This scheme limits a student's incorrect responses to previously identified alternative conceptions, so effective items can be created only after exhaustive interviews or through open-ended tests (Haladyna, 1994). The inclusion of an "I don't know" or "none of the above" category tends to reduce the efficacy such items because students are no longer obliged to choose among conceptions that may closely, but not exactly, match their own. They must judge, of the answers presented, which is best.

Multiple choice tests have been constructed, validated, and used in several different areas of science: light (Bouwens, 1986), mechanics (Halloun & Hestenes, 1985), cosmology (Lightman & Miller, 1989), electricity (Shipstone et al., 1987), chemistry (BouJaoude, 1992), and earth science (Schoon, 1988). These multiple choice tests produce standardized answers that can be compared. Such tests are easily scored even by those with no knowledge of alternative conceptions, and their results are useable immediately. They are useful not only for studies, but as diagnostic tools for teachers to ascertain alternative conceptions in their own classrooms.

Some argue vehemently that multiple choice tests have no role in the exploration of children's ideas, even choosing to dismiss any consideration of their value.[2] They are seen as part of the outdated technology that can test only trivial, low-level thinking (Haladyna, 1994). There is currently a preference for assessment methods that deal with fewer subjects in greater depth: concept mapping, clinical interviews, writing about demonstrations, journal study, etc. However, complex and expensive assessments are not always the best for every use (eyecharts outperform computed axial tomographic scans for the detection of nearsightedness). Over time, there is a trend for tools to become simpler and move out of the hands of experts (pregnancy tests were once accessible only through the medical establishment, and now, much to the relief of rabbits worldwide, have been supplanted by home tests).

Essay and equivalent multiple choice tests can be highly correlated, yet the multiple choice test is far more reliable (Haladyna, 1994). This is not to say that mental processes are easy to measure or that multiple choice tests of alternative conceptions are ready for use on standardized tests, but that we have an obligation as researchers to aid practitioners by preparing and studying tools that they will find valuable. Diagnostic tests based on student interviews and structured to be easy to administer, score, and interpret would be of great utility to science teachers.

Theories of Item Analysis

Multiple choice tests must be constructed, used, scrutinized, and refined. There are two major approaches to analyzing multiple choice tests. *Classical test theory* (CTT), the older and more straightforward method, provides tools for analyzing performance based directly upon the total score of a subject. *Difficulty* of test items is calculated simply as the fraction of students answering a question correctly. The ability of items to *discriminate* between students who do well on the test and those who do poorly can be calculated using correlations between single items and total scores (and other measures). CCT can be carried out by breaking the population into equal-sized groups (say two groups) based on total score and comparing the performance of each group on an item-by-item basis. CTT has historically been used in almost every quantitative analysis of multiple choice tests of students' ideas. Using classical test theory, results from a single population cannot be used to characterize populations that differ. Since student scores are test dependent, subjects undergoing different forms of assessment cannot be accurately compared.

Item response theory (IRT) was developed to surmount the apparent limitations of classical test theory in which measurements are highly dependent on a test population (Hambleton,

Swaminathan, & Rogers, 1991). IRT is a probabilistic measurement model based on the presumption that performance on a test is the result of only a few, and usually a single, ability or trait that cannot be measured directly. The total score on a test is related to but is not the same as this ability. The probability of choosing a particular answer on a test item can be expressed as a function of the subject's underlying ability and certain parameters of the item or question. Such an analysis treats the subjects as a sample from a larger population, breaking the continuum of performance down into a continuous spectrum rather than a series of discrete performance categories. The complete model bestows certain benefits:

- Subjects can answer different subsets of items and still be accurately compared.
- Item parameters are no longer dependent on a single population of subjects.
- Subjects of vastly differing ability can be accommodated on the same ability scale.

The purpose of IRT is to provide a basis for making predictions, estimates, or abilities measured by a test (Hambleton, 1989). Procedures for assessing the fit of the model to the data are well developed in the literature (Baker, 1985; Birnbaum, 1968; Hambleton et al., 1991).

Prior Studies of Alternative Conceptions Using IRT

Two previous studies have broken new ground by using IRT to analyze students' conceptions. The first examines mathematical alternative conceptions in a population of first-year college students (Narode, 1987). Items that diagnose algorithmic errors were developed for the mastery learning movement of the 1960s and 1970s. Narode found that items using the most common wrong answers as distractors were more difficult and their discriminating power lower than multiple choice math tests without such attractive answers. Narode also hypothesized that most standardized tests exclude items dealing with alternative conceptions because they do not fit the accepted item profile, as inclusion of such items would reduce the reliability of the test.

The second study fashioned a continuum of ideas from novice to expert, showing the development of student understanding about the particulate nature of matter (Doig & Adams, 1993). This was carried out by having 975 students respond to cartoon problems encompassing physical and chemical changes while making pancakes. Their comments were read and coded by raters who matched responses with qualitative categories generated from the alternative conceptions literature. A *partial credit model* (PCM) was used to find threshold values between these categories (Masters, 1988). A threshold represents the ability level at which the probability of choosing one model or answer increases above another. The authors found that their work resulted in a quantitative continuum "based on student data rather than only expert opinion" (Doig & Adams, 1993, p. 12). They also concluded that this single scale could be used to characterize students' conceptions across a variety of scenarios, stating, "whereas more common approaches would leave these facets as separate entities, our analysis provides a single, integrated view of students' understanding of all these facets" (Doig & Adams, 1993, p. 12). Doig and Adams posited that it is "possible to provide a continuum, rather than a collection of 'bits' . . . this approach has demonstrated the possibility of describing student performance across disparate but related notions" (Doig & Adams, 1993, p. 11.) This model grew out of attempts to apply IRT to scales with ordered categories allowing the calculation of numerical thresholds between each of these choices (Samejima, 1969).

There are two limitations to a PCM data analysis. First, it presupposes the existence of an ordered rating scale before such an analysis can be carried out. Second, when applied to multi-

ple choice tests, student conceptions that are popular but do not capture the majority of subjects at some ability level are not included in the model parameters. Attempts are currently under way to built proficiency scales without assumptions about ordering (Masters & Mislevy, 1993). This article is one such attempt.

There are alternatives to IRT for spreading a population out based on some ordered parameter. Whereas IRT uses an unmeasured, so-called "latent" variable, ability, one can use a more concrete variable, such as age. A systematic study of predictions of mechanical motion was carried out with 631 students aged 7–18, using an open-ended test (Eckstein & Shemesh, 1993). The predictions were of projectile motion on earth and in space, relative motion, and circular impetus. The authors fit a mathematical model to their data. Their premise was that students move through three stages on their path to scientific understanding. The authors found that the model was consistent with their data, although reduced chi-square statistics never had a probability $>.9$ for repeated measures. One reason is that the spread in scores by age is typically much larger than the change in score between ages. Another is that the authors chose three stages and fit all students to those only, despite the inevitable variety of qualitative responses. A third problem is that with only three parameters per item, the shapes of the trace lines were highly constrained. IRT solves all these difficulties by constructing an ability scale based on all questions, by employing any number of alternatives and estimating a latent "don't know" category, and by modeling each item response with one to three variables. This categorical response model (Bock, 1972) results in a superior fit to the data with fewer assumptions about hierarchical order or the rates of passage from one stage to another. Despite its more abstract nature, IRT is more suited to generalizing to larger populations and to characterizing item responses.

Method

The body of this article is devoted to the analysis of a specially constructed instrument, the Project STAR Astronomy Concept Inventory. It is a 47-item multiple choice test designed as an assessment instrument for a National Science Foundation–supported curriculum project.[3]

Subjects

The population used for the analysis in this article totaled 2562 responses of which 1250 students took both a pretest and a posttest. The remaining 62 post–high school subjects took it only once (Table 1). Designed as a norm-referenced test to compare the effect of different instructional treatments, it was used over a 5-year period with many thousands of students comparing the effect of year-to-year modifications in curriculum materials. By collecting data from astronomy and earth science classes that were either using Project STAR or other curriculum materials, we were able to measure the impact of a novel hands-on curriculum which explicitly deals with student alternative conceptions in astronomy.

The subjects are students from 22 classrooms, of which 19 were testing Project STAR materials (7 in Boston, 12 nationally). Three classrooms were selected in the same schools as Project STAR teachers and two more from other schools. Three of the schools were in rural areas, 16 in suburban districts, and 3 in cities. All but two were public schools. The economic status of these communities varied considerably. Four were characterized by these teachers as low, five as low to average, nine as average, two as average to high, and one as high. School sizes varied from 325 to 2000 students, with a mean of 1282.

Table 1
Population used for IRT model

Group	Mean Pretest Score (Out of 47)	SD	Pretest	Posttest	% Pop
Grade 8	13.63	4.91	344	Yes	13
9	15.18	6.44	156	Yes	6
10	16.15	5.40	121	Yes	5
11	17.03	6.19	275	Yes	11
12	18.65	6.93	267	Yes	10
Unidentified (8–12)	15.59	8.01	87	Yes	3
Undergraduates	25.38	4.40	20	No	1
Astronomy majors	41.44	3.10	9	No	1
Astronomy grad	43.00	2.75	15	No	1
Astronomy professors	44.11	3.05	18	No	1
Total			1312	1250	

Note. The model was built from treatment and control groups for Project STAR (1250 from Grades 8–12) to which Harvard students and professors were added to calibrate the high end of the scale (62 additional subjects). Note that the yearly increase in mean score is much smaller than the standard deviation in scores within a grade.

Instrument

The content, construct, and criterion validity of the test were investigated in depth (Messnick, 1989). Astronomy teachers were asked to predict the percentage of students correctly answering subsets of items on the test as a measure of how well this test dealt with the concepts they taught in their own classes. As a group, the teachers predicted that only 36% of the questions would be answered correctly on the pretest and that 73% would be answered correctly on the posttest. The predicted gains support the claim that teachers thought the test was a reasonable measure of many of their course objectives.

In a nationwide survey, 240 astronomy teachers predicted the pretest and posttest scores of their students on 16 selected questions from the Project STAR inventory (Lightman & Sadler, 1993). These questions were chosen as having the greatest similarity to the concepts taught in introductory astronomy classes. Graduate students in Harvard's Department of Astronomy took a version of the inventory, pointing out inconsistencies or other problems with items. Astronomy teachers took versions of the test to see if they chose the correct answers. Item characteristics were explored to seek plausible explanations of why students chose not to answer certain questions.

Procedures

The Project STAR curriculum aims to help students construct their own astronomical knowledge. The course focuses on three major concepts in astronomy and their models: physical systems of various sizes, the behavior of light, and the use of simple mathematics to solve scientific problems. Test items were generated to match these objectives in two ways: by searching the literature for studies of children's alternative conceptions in astronomy, and through interviews and open-ended tests of students about their ideas by project staff[4] and consulting teachers.[5] Each test item consisted of a stem and five alternatives. Only one of the responses

was scientifically correct; the other four represented alternative conceptions. These *distractors* were written to be as plausible as possible. An effort was made to ensure that all answers were similar in length and complexity and contained a minimum of scientific jargon. Additional information on this test and its development can be found in Sadler (1992).

We began our curriculum project by interviewing students in high school earth science and astronomy courses. It quickly became apparent that students were not simply discarding their prior conceptions for scientific ones. For some, studying astronomy appeared to have strengthened their prior ideas. Others had abandoned their views for alternative conceptions that they found more appealing. Still other students had proceeded from not having any firm convictions at all to firmly held alternative conceptions. A study of the pretest/posttest contingency tables for the Project STAR questionnaire supported the theory that most students were not progressing steadily to scientific conceptions, but taking conceptual detours. It seemed reasonable to try using the test as a tool to help place students on a performance continuum. For this criterion-referenced interpretation, CTT statistics are not appropriate and IRT methods were investigated instead. After completing the testing of 1250 students in the final year of the project, there were none with a raw score above 40/47. As an aid to understanding how the test behaves at high performance levels and to keep the analysis accurate at high ability levels (Haladyna, 1994), 75 Harvard students and professors were invited to take the test. The use of such expert judges also increases the face validity of the test.

The IRT Model

One of the features of test theories is the ability to spread the population into a spectrum of abilities (based on total score) and graph the resulting response patterns. These have been termed *item option characteristic curves* (IOCC). Using this technique, several characteristics of the resulting trace line become apparent. The probability of choosing a correct answer usually rises with ability. The chance of choosing a distractor generally decreases with ability. The difficulty of items can be characterized by where the correct answer rises above a threshold probability of 0.50 (Hambleton, 1989). IRT models differ from CTT models in that the population is not spread into a discrete number of bins, but the data are fit to a logistic distribution function. The horizontal axis, termed "θ," or ability, is generated in the process of curve fitting and is related to the raw score, but is transformed, stretched, or shrunk to best fit the data for the entire test. This same θ is fit simultaneously to every item. Theta scores for different subtests will be identical.

Many IRT models have been investigated for this study, involving one to three parameters and variants on logistic and other functions. For the purpose of studying children's alternative conceptions a three-parameter, nonmonotonic model of this form was used, developed by Bock in 1972:

$$P(x_j = k;\theta,a,c,d) = \frac{e^{z_k} + d_n e^{z_0}}{\sum_{h=0}^{m_j} e^{z_k}} \tag{1}$$

where x_j is the response to the item and $k = 1, 2 \ldots m_j$ for multiple choice item $_j$ with m choices. θ = the ability; a, c, and d = the item parameters; $z_0 = a_0\theta + c_0$ = a linear function of θ for the "don't know" category; and $z_k = a_k\theta + c_k$, a linear function of θ for each of the item responses.

This model has many interesting features. Since we are intent on understanding student selection of the item distractors, this model incorporates each of the distractors along with the correct answers. Most IRT analyses parsimoniously examine only correct responses. Including distractors has been found to generally improve accuracy for subjects in the lower half of the ability range (Hutchinson, 1991). By using a function that can also be nonmonotonic, one can determine the degree to which alternative conceptions may be of increasing attractiveness to students of higher ability. This is a critical issue. Although student alternative conceptions are generally thought of as becoming less attractive to older students or to those who have more knowledge, there have been examples in the literature (albeit with small numbers of students) that show alternative conceptions appearing to rise with age and ability before they decrease.[6] The three-parameter model also has the advantage of estimating student guessing as a function of ability (the "don't know" category) without placing such an option directly within each test item (Thissen, Steinberg, & Fitzpatrick, 1989). This category is not an observed response but a latent one, and has the benefit of estimating the proportion of students who guess at each response from those who choose answers intentionally (Thissen & Steinberg, 1984). This bestows an advantage, since the "I don't know" option is chosen differentially by students of differing genders and ethnic backgrounds, creating test bias by benefiting some test takers over others (Haertel & Wiley, 1993; Haladyna, 1994; Sherman, 1976). Item parameters were calculated using the computer program Multilog (version 6.0, 1991) for a PC-based platform. Multilog performs multiple, categorical item analysis and test scoring using IRT.[7]

Results

It is not possible within the scope of this article to present a full analysis of each of the 47 items in the Project STAR questionnaire using IRT. Instead, three representative questions from the test will be discussed in detail and followed with two different methods summarizing the characteristics of all the test items.

Three Representative Items

The Reason for Day and Night. The reason for day and night is perhaps the most basic idea assumed by teachers of astronomy of their introductory students. In a survey of 330 secondary school earth science and astronomy teachers, participants predicted that 65% of their students, on average, would enter their classes with this concept understood, and by the end of the course 89% would leave knowing it (Lightman & Sadler, 1993). The earth turning on its axis causing day and night has been described as one of the "most essential ideas which form the Earth conception" (Nussbaum, 1985).

The multiple choice item used for examining this idea was constructed by carefully identifying research in this area and using the results from several qualitative studies of students' conceptions of day and night. For example, 20 students were interviewed (Grades 3–10), resulting in two types of explanations of the day/night cycle: The sun goes behind something (hills, clouds, or the moon) or something spins (the earth, the sun around the earth, or the earth around the sun) (Baxter, 1989). Further work pointed to the gradual replacement of the view that the earth goes around the sun with a spinning earth. In a study of 60 elementary school students (Grades 1, 3, and 5), the reasons stated for day and night included that the earth spins, the earth revolves around the sun, the earth or sun moves into a shadow, clouds block out the sun's light, or "God made it that way" (Vosniadou & Brewer, 1987).[8]

Children as old as 12 years believe that the world is a spherical shell with the earth comprising the bottom of the hemisphere, and air in the top half. The sun travels along the surface of the sphere in this model. Children think that the sun is in the sky in the daytime and travels below the earth at night (Nussbaum, 1985). Another 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 moves (Klein, 1982). Item 1 was written by selecting the most common answers found by researchers in the above studies. Since this test is designed for Grades 8 and up, conceptions that appeared to be unique to younger children (e.g., God made it that way) were not used.

Item 1, Reason for Day and Night.

What causes night and day?

A. The earth spins on its axis. (0.66)
B. The earth moves around the sun. (0.26)
C. Clouds block out the sun's light. (0.00)
D. The earth moves into and out of the sun's shadow. (0.03)
E. The sun goes around the earth. (0.04)

In our population, the majority of subjects chose the correct answer (underlined). Classical test theory assigns this item a difficulty of .66. It is easy to imagine that this parameter would be different with another population (e.g., Harvard professors scored 1.00).

The probability of choosing each answer in Item 1 is depicted graphically as a function of performance on the entire test (Figure 1). The horizontal axis is the overall ability score in units of standard deviation from the mean population score. The vertical axis is the probability of a student of a particular ability selecting a particular answer. The guessing estimate calculated by the model represents the proportion of students who do not discriminate between the answers for this item.

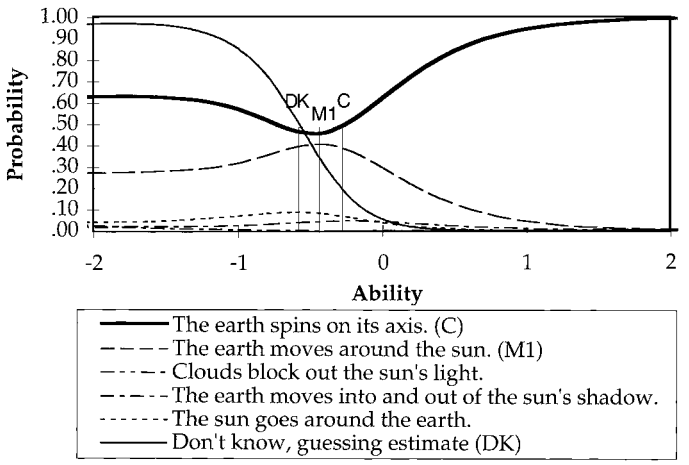

Figure 1. Item option characteristic curve for Item 1: the cause of day and night. The probability of selecting the correct answer (C), that the earth spins, is lower in students of moderate abilities than those of low abilities or high abilities. This corresponds to an increase in preference for the incorrect notion of the earth's orbit being responsible (M1). The estimate of guessing (DK) falls dramatically as ability increases in the population.

The generally accepted profiles for these trace lines are that the correct answer should monotonically increase and distractors should monotonically decrease (Haladyna, 1994). These curves do not match the nonmonotonic behavior of traditional tests. The probability of selecting the correct answer is 0.63 with low-ability students (for $\theta = -2.0$). With moderate-ability students, the probability decreases to a level of 0.46 before it rises in high-ability students to 1.00. In IRT, item difficulty is defined as the ability level at which the correct answer rises to 0.50 (Hambleton, 1989) and is marked *C*. The difficulty of this item is −0.2. The most popular distractor, that the earth moves around the sun, is also nonmonotonic (marked *M*). It rises in probability before it descends, peaking at an ability of −0.4. The estimate of guessing, the "don't know" trace line, drops from 0.97 to zero as ability increases. It estimates that half of students have "no idea," but may still guess correctly, at the −0.6 ability level (Thissen & Steinberg, 1984) and is marked *DK*.

It is important to recognize that the model shows that the population exhibits a spread in the ability levels at which guessing drops to 50%, the most popular distractor peaks in appeal, and the correct answer rises to 50%. This spread is the distance between the *DK* and *C* vertical lines. The misconception *M* is marked as vertical line within this range. It is reasonable to assume that overall performance on the test increases with course taking in science and with age (this is borne out later in the article). The pattern of increasing ability in the population from "don't know" to popular distractor to scientific answer may model the growth of individual students.

The Reason for Seasons. The second question analyzed has to do with the cause of the seasons. The sun appears lower in the sky during the winter than in the summer. This change in altitude spreads the sun's light over a much broader area on the earth's surface. This fact was well known to the Greeks and works equally well in either heliocentric or geocentric cosmologies. That the earth orbits the sun is of no special consequence to seasonal change; the sun orbiting the earth would produce the same result. The constant angle that the earth's spin axis makes with the plane of its orbit (from a heliocentric perspective) or the yearly cycle of movement of the sun along the ecliptic (in the geocentric model) produces the same effect on earth, the sun's seasonal change in altitude.

The reason for seasons is almost universally taught in the fifth or the sixth grade. The Boston Curriculum Objectives (Marshall & Lancaster, 1983) for fifth grade explain correctly that the reason for winter is that "the sun is lower in the sky," but then go on to qualify this reason with the questionable statement, "Its rays have to shine through more atmosphere before they reach us, losing heat energy in the process." Many students believe that the change of seasons is evidence of the earth's elliptical path (Touger, 1985). Others explain that the earth is simply closer to the sun in the summer than the winter, without reference to an orbit (Furuness & Cohen, 1989). In 1986, three Italian researchers found that the majority of 11- to 13-year-old pupils believed that "the sun always rises from the same point on the horizon, the east, and always sets in the opposite point, the west" (Loria et al., 1986). Schoon found that 12 of 13 participating teachers and 20 of 32 student teachers believed that the sun is always overhead at noon (Schoon, 1988), so there would be no effect owing to solar insolation changes (change in altitude).

Item 17, The Reason for Seasons.

The main reason for its being hotter in summer than in winter is:

A. The earth's distance from the sun changes. (0.45)
B. The sun is higher in the sky. (0.12)

C. The distance between the northern hemisphere and the sun changes. (0.36)
D. Ocean currents carry warm water north. (0.03)
E. An increase occurs in "greenhouse" gases. (0.03)

The scientifically correct answer is chosen only 12% of the time in the Grade 8–12 student population, less frequently than if chosen at random. It is answered correctly 100% of the time by astronomy professors and graduate students. This is a question that is much more difficult than Item 1. Teachers in the prediction survey thought students in introductory astronomy courses would score 0.29 at the start of their courses and 0.76 by the end. Distractors A and C are far more popular than the scientifically correct response in the student population.

In answering this question, students appear to be torn between two distractors that mention changing distance. Many students believe that the earth's orbit is highly eccentric so that the entire earth is physically closer to the sun in the summer than in the winter (Newman & Morrison, 1993; Sadler, 1987). A more evolved explanation is that the earth leans toward the sun in the summer and away from the sun in the winter, accounting for the difference in seasons in the northern and southern hemispheres. This is consistent with many diagrams in textbooks that show the one pole proportionally much closer to the sun in the summer. This fact is an indicator, but not for the cause of the seasons. It is the spherical shape of the earth that results in seasonal variation across latitudes. Were the earth shaped like a rolling pin, all latitudes (but the flattened ends) would have the same seasons simultaneously.

Again, these curves do not match the nonmonotonic profile of traditional items (Figure 2). The "don't know" curve drops to a 0.50 level at a θ of −2.4. The most popular distractor peaks at a θ of −0.1. The second most popular distractor peaks at a θ of 1.2. The correct answer attains a probability of 0.50 at a θ of 1.7. This profile is different from that of Item 1 in several ways. Two distractors appear to be very popular compared to only one in Item 1. The probability of choosing these two distractors rises above all the other trace lines for students of mod-

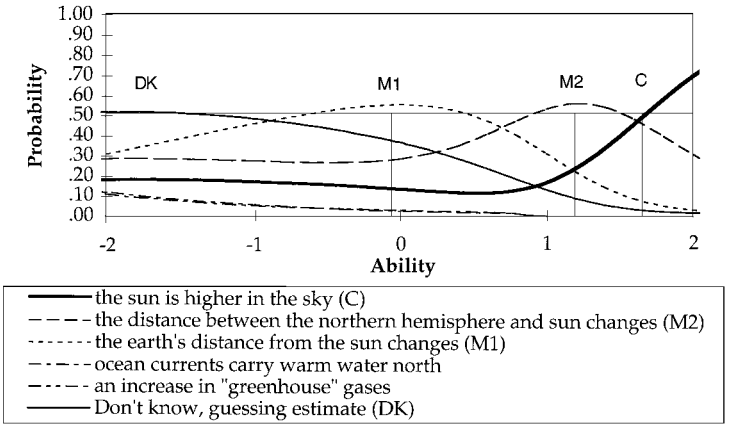

Figure 2. Item option characteristic curve for Item 17: the reason for seasons. The preference for the correct answer (C) declines in the population until a very high ability is reached (0.4), after which it rises dramatically. The preference for wrong answers is dominated by two conceptions (M1 and M2), both mistakenly involving distance from the sun as a factor in changing seasons.

erate ability. In a group of students with an ability of 0.0, the distance model for seasons would be the most popular of all the choices for this question. A group of students at the 1.2 ability level would find the hemispheric distance difference the most plausible answer. The responses to this item in the student population suggest a progression in thinking about the earth's seasons from not discriminating between answers (don't know) to a solar distance model to a hemispheric difference model to solar altitude model.

A Model for the Sun and a Close Star. The third item analyzed in detail examined the mental models that students construct for the relative disposition of stars. Stars are very far from us on the scale of the solar system. The closest star to our solar system is about 200,000 times farther from it than the sun is from the earth. Or, to put it another way, if the sun were a grape the earth would be a speck of dust 3 ft away and the closest star would be another grape 100 miles from us. At this scale, even Pluto would be close to the sun, at a distance of 100 ft. Item 8 asks just this question with five options for answers; two grapes would make a good scale model of the sun and a close star, if separated by 1 ft, 1 yard, 100 yards, 1 mile, 100 miles. Since stars and planets are almost indistinguishable in the night sky, students may not make a distinction between their distances from us. Among 200 11- to 13-year-old Italian students interviewed, there was no distinction between stars and planets (Loria et al., 1986).

The probability of answering this question correctly declined slightly from −2.0 to 0.0 and then rose sharply, reaching a value of 0.5 at an ability of 0.9 (Figure 3). The "don't know" option dropped to a value of 0.5 at a θ of 0.2. Examining the detail of the trace lines of the four incorrect options shows a progression of preferences from 1 ft at ability −0.2, to 1 yard at ability 0.1, to 100 yards at ability 0.7, to 1 mile at ability level 0.8.

This result is quite surprising in that it was thought that the item was testing for knowledge of a fact, a memorizable number. Yet, student choice of distractors followed a pattern that con-

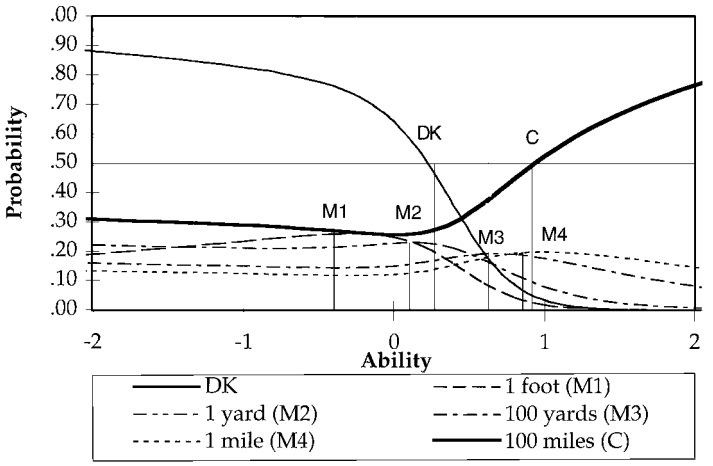

Figure 3. Model of the sun and a close star as two grapes. The J-shaped trace line for the correct answer shows a lower preference among midlevel students than low-level students. This dip corresponds to a preference of answers that correspond to models that put stars increasingly further from each other. Each of these preferences has a corresponding ability level which can be compared to other test items.

verged on the correct answer with increasing accuracy, from 1 ft to 100 miles. All professors selected the correct answers; 0.68 of graduates and undergraduates answered correctly, with the remainder choosing "1 mile" (M4). The "fact" of the scale distance between two stars is tied to the student's model of celestial objects; it is not thought of in isolation. In diSessa's terms, the model of celestial objects in space is the phenomenological primitive for notions of individual sizes and distances. This is supported by the multitude of misunderstandings that come from thinking astronomical objects are relatively close to each other. These include the differential model for the seasons and lunar phases being caused by the shadow of the earth.

Goodness of Fit

How well does this IRT model fit the data? Unlike classical test theory, there is no single measure of goodness of fit. The fit of IRT models must be tested using a variety of methods (Hambleton et al., 1991). A CTT analysis was first conducted (Sadler, 1992). There was a large variation in discriminating power [mean = 0.29, standard deviation (SD) = 0.16] of the items (as measured by item: total test correlations). The high average difficulty of the items called for a three-parameter IRT model. A principal component analysis revealed that a dominant first factor was present. The magnitude of the first eigenvalue was 3.4 times the magnitude of the second, and the second largest was barely distinguishable from the rest. This is evidence of unidimensionality (Hambleton, 1989). The invariance of the item parameters was demonstrated by calculating them for different subgroups in the population, including pretest and posttest scores, and control and treatment groups. The invariance of ability estimates was tested by both comparing parameters for different samples of test items and by calculating test characteristic curves for different sets of items. The fraction of subjects choosing each answer was compared with that calculated from the IRT model for each of the 47 items. The SD of all such errors was 0.011.

The contribution that each item makes to an estimate of a subject's ability depends on its IOCC. Steep portions of curves contribute most to the ability to discriminate ability; flat portions contribute least. A standard error of estimation of ability for the entire test can be calculated from the IOCC for each item. Using only correct answers and ignoring the contribution of distractors, the standard error of measurement of ability is within a reasonable range, 0.30 SD, from a θ of -0.2 to 2.5. The inclusion in the model of all distractors had the effect of extending the range of high precision in ability estimation to a much lower ability level of a θ of -0.9 to 2.5 and increasing the information in the test results (Thissen & Steinberg, 1984). This matches well with the high school population studied. The θs for 90% of the students are estimated with high precision. Only those scoring in the lowest 10% (fewer than 10 items correct) were accounted for with lower precision than the rest of the population.

Combining and Comparing Test Items

The test items described above are to a large degree representative of all the items on the test. All items have nonmonotonic trace lines for at least one distractor; almost all exhibit nonmonotonic trace lines for the correct answer. This is highly unusual. Almost every test item reported in the literature on standardized tests is monotonic for both correct answers and distractors. This behavior suggests that the learning of scientific concepts may not best be modeled by the slow accumulation of knowledge over time, but by more complex, nonlinear processes. It is worthwhile to examine ways of looking at the data to determine how this test differs from others and to develop ways of looking at the meaning of the θ peaks for the distractors. U-shaped learning curves were documented in the literature concerned with knowledge states dependent

on rules, such as learning verb endings (*walks* and *walked* leads to *eats* and *eated*) (Richards & Siegler, 1982).

All 47 items had hump-shaped trace lines for at least a single distractor. Of a total of 47 items on this test, 45 had J-shaped trace lines for the correct answer. Such nonmonotonic trace lines have been revealed by other IRT researchers and commented on extensively in the literature (Choppin, 1985), and generally interpreted as evidence for cheating. This test was low stakes; scores had no influence on grades so that there was no motivation for cheating by students. Hutchinson (1991) summarized the arguments for and against using such questions on tests. Those who thought that such items are a bad idea used arguments that such items have the "character of a riddle," include "harmful distractors," or "should not be used unless there is a some special reason," or that these are questions of "tricky character." Few were more supportive, stating "this is a desirable feature" or that such distractors "[predict] the kinds of misunderstandings or errors in knowledge that examinees are likely to experience."

Psychometricians construct a test characteristic curve (TCC) by summing the correct IOCC for all test items. This gives a general picture of the difficulty and discriminating power of the combined set of test items. It also provides a way to transform ability scores to test scores (and back again). Figure 4 shows the TCC for the Project STAR Questionnaire. It is relatively flat in the ability range of −1.0 to −0.5, rising steeply to 2.0, and then rising gradually from there on. Interpreting individual items can be problematic in that some students appear to be penalized should they answer correctly.[9] This would raise their ability estimate more than choosing the correct answer. When summed over the entire test, the bumpy shape of these individual items appears to average out, looking much like TCCs from more conventionally constructed tests. The curve representing correct answers takes the form of an almost monotonically increasing function and is independent of the distribution of the population in ability (Baker, 1985).

Figure 4 shows a plot of the mean scores of the various groups that took the test. It should be read against the horizontal ability scale. This shows the power of an IRT analysis in its in-

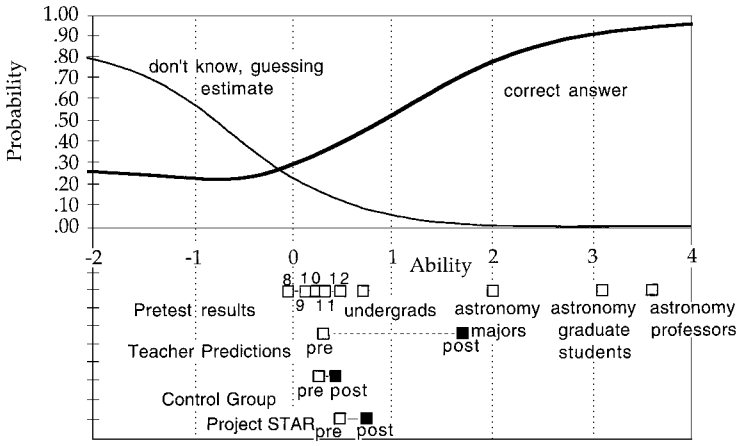

Figure 4. Test characteristic curve and population means. The overall test performance of groups is represented by their horizontal position of populations on the lower portion of the graph. One can tell that teacher pretest predictions appear to be close to the actual scores of the control group. Ability levels appear to grow with increasing grade level of the population. The correct answer trace line is the average of all items on the test. It show a slight dip with ability and rises steeply in the range of abilities from 0.0 to 2.0.

clusivity of groups with a wide range of ability. One can see the mean pretest score of the high school students (Grades 8–12) in the range of −0.1 to 0.4, as well as a sample of Harvard undergraduates taking an introductory astronomy course, astronomy concentrators, graduate students, and astronomy professors. The pre- and posttest results of a control group taking high school earth science or astronomy is compared with the results from the group of Project STAR students. Teacher predictions for 16 items from the test are compared to those for the total test. This scale can be used to compare groups. Starting with the pretest results, one can see a small but measurable difference in the ability levels of students entering astronomy or earth science courses. Those in higher grades are at higher ability levels. Since students have not yet taken an astronomy course, this difference can best be thought of as representing the natural learning that takes place from year to year without the help of a specialized course. The average yearly gain is 0.13 SD in ability.

In contrast to this natural gain, students in conventional astronomy courses increased an average of 0.16 SD, a larger change than students presumably would have achieved without the course. Students in the Project STAR course increased 0.27 SD, more than twice the gain of students without the course and nearly twice the gain seen in conventional classes.

It is useful to compare observed ability levels with those predicted by teachers. Teacher predictions were remarkably accurate for the start of the course. However, they fell outside of the range of students' final performance, rising to an ability mean of 1.70 for the end of the course (Figure 4). This is close to the average performance of undergraduate astronomy majors at Harvard University. It requires an average gain in ability of 1.40 SD, roughly eight times the gain seen in traditional astronomy courses and five times the gain seen in Project STAR classrooms. Teachers predict that they are helping students move to a much higher level of understanding than they actually attain.

The IRT model can now be harnessed to explore the reason for this overestimation, since the model spreads the population into a continuum of ability that is shared by each test question. One can now analyze individual test items to find out how each group performed, i.e., what was their most probable answer? What is the trend by grade for each answer? How accurate are teacher predictions for gain resulting from a single year of instruction in astronomy? How do the ability levels of different groups compare?

Returning to Item 1 (Figure 1), we can now examine student responses as a group in Grades 8–12 (ability level −0.1 to 0.4). In this range, the IOCC for the correct answer and for distractors was unremarkable. It followed traditional expectations for monotonicity. Answers suggestive of alternative conceptions probably peaked at much lower ability levels, perhaps within grade school. Administering this test at these grade levels could help to identify the grades in which these alternative conceptions peak. As a teacher, one would expect any alternative conceptions held about the reason for day and night to decrease in popularity over a year of astronomy or earth science. The incidence of alternative conceptions in older children, high school seniors, and college students would be rare.

The cause of the seasons (Item 17) is more revealing. For students in Grades 8–12, the belief that the earth's changing distance from the sun is responsible for the seasons was very popular and near its peak in popularity. A year of science appeared to do little to change this belief in the sample. Moreover, Figure 2 reveals that those students who did abandon the elliptical orbit theory probably shifted to another alternative conception, coming to believe that the differential in distance between northern and southern hemispheres is responsible for seasons. Following the trace line for the correct answer, one can see that the number of students who chose the correct answer declined for the high school population, whereas teachers expected that students would abandon their alternative conceptions by the end of their courses. This expectation

was quite optimistic; the cause of the seasons evidently took a very long time to learn. Judging from Figure 2, to move from the initial conception to a correct understanding may very well take 8 years of study.

Students' spatial model of our starry neighbors appeared to change somewhat in high school (Figure 3). Although the majority of students did not appear to prefer any single distance between stars (revealed by the "don't know" trace line), a pattern could be seen in that longer and longer distances were preferred by students as they aged. However, the portion of the IOCC for the correct answer with the steepest slope occurred at ability levels higher than high school.

The analysis above presents a picture of learning that is not steady or uniform. Such a finding is not new, but the evidence here is quantitative and substantial. It has long been recognized that children can acquire new alternative conceptions as a result of instruction (Linn, 1986). A longitudinal study of children's conceptions of conservation of mass during combustion found that students changed their view that burning would decrease mass in a closed system to that of burning increasing the mass of the system (BouJaoude, 1992). Understanding of chemical changes can also progress from "magical" explanations to those that only recognize directly observable processes (Doig & Adams, 1993). A qualitative study found a child's understanding of the changing seasons began with believing the act of adjusting a clock from daylight savings to standard time accounted for the changing duration of daylight. Later in the year, the same student explained the seasons with a changing distance model (Michaels, 1995). What this study uniquely reveals is that such shifts often are markers of progress, even though such conceptions appear equally wrong or confused to experts. Moreover, children appear to regress in measures of their understanding of certain concepts while they make progress on others. While the trace lines for correct and incorrect items were bumpy for the items discussed above, the TCC was smooth and monotonically increasing in the range for which the test was most valid, -0.9 to 2.5.

Representations for Multiple Items

Much can be learned by comparing the IOCCs for test items. Three items above were analyzed using IRT; the same analysis was performed on every item on the Project STAR test. From the IRT model, two to six parameters characterized each item: the θ values that are indicative of:

- the 0.50 level for the "don't know" option,
- the 0.50 level for the scientifically correct answer, and
- the ability level at which nonmonotonic distractors peak.

These values can be similar or very different. The range in θ between the first two values defines a special domain in which the majority of students neither are guessing among answers nor have settled upon the scientific explanation. This represents the span in which the process of shifting between competing conceptions takes place in the population, and it can be parameterized by its θ end points and by the range between them. If this range is small (as it is for the reason for day and night), it can be argued that the shift from "don't know" to correct will take place relatively rapidly (perhaps within a few years of schooling or less). If this range is large (as it is for the reason for seasons), the cognitive shift will probably take far longer and be mediated by several stages in acceptance and abandonment of alternative conceptions.

This cognitive progression of conceptions representing the IRT results can also be used to match student populations with the desired learning outcomes. Eighth-grade students in our sample have a mean increase in ability from -0.05 to $+0.12$. Teachers can expect some suc-

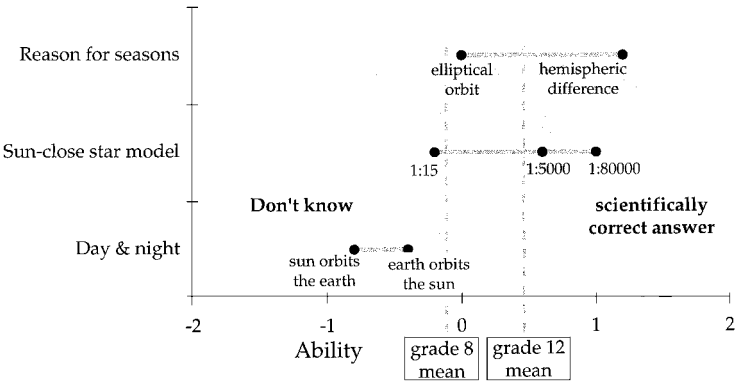

Figure 5. Cognitive range of three items. This chart is constructed from the item option characteristic curves of each included item. The gray area represents the extent of difference between the guessing dominating and the correct answer dominating for each item. Items are arranged in order of increasing difficulty. The dots mark the ability levels corresponding to the peak of each wrong answer and are connected together horizontally by item. Vertical lines mark the ability means for 8th- and 12th-grade students.

cess in getting almost all students to understand relatively easy concepts (the cause of day and night) for which the majority of students are already convinced. However, for those concepts marked by greater difficulty and a larger cognitive range, instruction may simply reinforce prior alternative conceptions or move students to accept other alternative notions (the sun-close star model, the cause of seasons). Older students and those of higher ability (Grades 11–12, introductory college courses) may be able to master these more difficult ideas. Of course, one must take care not to characterize students only by the mean ability of their classmates. Within each group there will be students far above and below this level. With excellent teaching, students may be able to reach higher levels of achievement. These ways of looking at data help to characterize only the average movement in a population, not an individual's precise path. For this goal, only longitudinal studies will suffice. Cross-sectional analysis helps by generating theories which can be explored in greater detail in future work.

It would be time-consuming and fruitless to discuss all of the responses to items in great detail. Instead, we can compare the cognitive ranges of the three test items we have already considered by representing the results in graphical form (Figure 5). The items are all plotted vertically by increasing item difficulty (in terms of ability). Like IOCCs, these graphs all share the same ability axis and one can move between graphs for comparison.

The graphs have two major components. The solid gray area is bounded on the left by the ability level at which the fraction of students who do not discriminate between answers (the "don't know" construct) drops to 0.50 of the population. On the right; the gray area is bounded at the ability level at which the fraction of students selecting the scientifically correct answer rises to 0.50. The gray area defines the range in which a majority of students discriminate between answers, but do not prefer the correct response. This domain represents a range in which students are "in limbo," having a definite preference that is not correct. The black dots on the graph mark the peaks on the ability scale for various alternative conceptions. If there are two or more, they are connected by horizontal lines to make them more visible. The two vertical dotted lines mark the pretest means for 8th-grade and 12th-grade students.

The concepts in this set of items are suggested for coverage in Grades 6–8 in the National Science Teachers Association's Scope Sequence and Coordination Project (NSTA's SS&C) (Aldridge, 1995). The mastery of an easy concept (reason for day and night) can be seen far to the left of the eighth-grade mean, but the rest of the concepts are not understood by the majority of high school students. The average 12th grader believes that close stars are thousands of solar diameters away (but not millions), and that the seasons are caused by some distance-related effect. These two concepts appear to be so difficult that the majority of students will have trouble mastering them in a single year of conventional instruction.

For the entire 47-item test, only the very easiest of items appear to have small cognitive ranges. As one climbs in difficulty, the span of fixation on alternative conceptions stretches to many years. Comprehension of vast astronomical scales appears to remain beyond the reach of students even after taking an earth science course of astronomy course in high school. Yet, modeling stellar distances is recommended for study in Grades 11–12 (Aldridge, 1995). What may be more beneficial for students at this age attempting to understand such concepts is *not* explicitly learning the scientific model, but gathering evidence for their beliefs and then being dissuaded from the more primitive model that they hold.[10] This strategy may aid in the transition to more sophisticated alternative conceptions or the scientifically correct explanation.

It is also instructive to examine a set of closely related concepts. The moon's motions and interaction with the earth are typically taught in Grades 6–8, as are recommended in the new national standards (Aldridge, 1995; Council, 1996; Project 2061, 1993). Most students have not mastered the facts of the moon's orbital period by Grade 8, but by Grade 12 they have learned them (Figure 6). However, most 12th graders in this study still cannot recognize a scale earth–moon model and its component distances. This model is the critical logical element in refuting the role of the earth's shadow or sun's shadow in the moon's phases. It appears from Figure 6 that understanding and using the reason for the moon's phases are much too difficult for the majority of students even at the 12th-grade level, probably because they have not yet come to understand that it can only be the moon's own shadow cast upon itself which is responsible, and not the earth's shadow. Harvard astronomy majors were the lowest-level group tested for which a majority had mastered this concept.

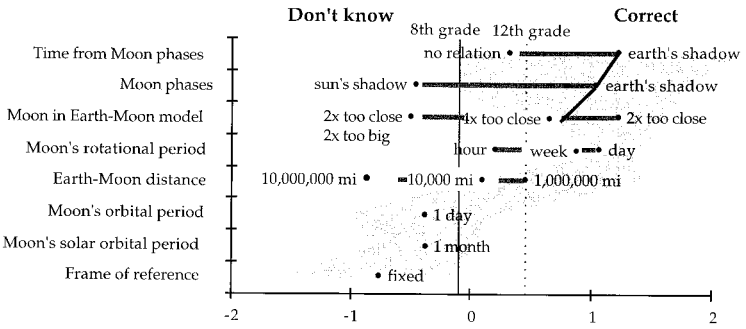

Figure 6. Cognitive range of earth–moon items. The lowest two items appear to be answered correctly by many eighth graders, while the five most difficult questions appear to be too difficult for 12th-grade students. For each item there is a least one distractor which peaks and is preferred by many students. For example, the earth–moon model preference appears to change from similarly sized objects that are close together, to correctly sized object spaced too close to a correct model. This spread stretches over a range of ability that corresponds to a time period much larger than the 4 years of high school.

The similarity of three item choices—the earth's shadow being used to explain sunset time, the earth's shadow explaining phases, and an extremely close earth–moon distance—are again indicative of a belief that objects interact only when they are close by (connected by a black line in Figure 6). This is likely the model of an object falling in the shadows of other objects. Correct in itself, it is not the explanation for the moon's phases. Again, a study of a scale model of the system would do much to dissuade students from their view, much more than providing an explicit description of the scientific model. Moreover, the most common demonstration of the moon's phases (using a nonscale lightbulb and ball) offers only an alternative model for the moon's phases. A scale-model demonstration is essential to show that the earth's shadow *cannot* have a role in phases.[11] This concept may be so difficult that only college-level students can have success in learning it.

These two multiple-test item analyses have similarities. While IRT analysis using distractors lays out the relative ability level associated with each answer, the two multiple item graphs (Figures 5 and 6) allow for comparisons between items. These representations also allow comparison of the levels of items with the levels of students in different grades and populations. Every one of these concepts is thought by teachers to be learned in high school earth science and astronomy courses (as revealed in prediction studies). Yet, they vary widely in difficulty and in the level of prior alternative conceptions.

Discussion

The analyses generated previously, IOCC for individual items, a TCC, alignment of population pre- and posttest scores with the TCC, and multiple-item cognitive range graphs, all help build a quantitative picture of learning in astronomy.

Curricular Issues

Astronomy is a popular topic throughout the precollege curriculum. In elementary school, students first are exposed to seasons, moon, phases, and the planets. In junior high school, students begin their study of the solar system and stars, usually as a part of an earth science course. These courses serve an estimated 1,000,000 students each year (Welch, Harris, & Anderson, 1984). Somewhere between 10% and 15% of U.S. high schools offer semester or yearlong elective courses in astronomy (Sadler, 1992; Weiss, 1987). Introductory college courses in astronomy are popular and often convenient ways to fulfill science requirements.

Astronomical concepts are an important part of efforts to construct national standards for science education. These efforts emphasize the historical development of the heliocentric system from Ptolemy to Kepler (Project 2061, 1993), the origin and evolution of the universe (Council, 1966), and energy in earth and astronomical systems (Aldridge, 1995).

Analysis of IOCC for each item shows that for apparently basic concepts, students prefer alternative conceptions that are often spread over a wide ability level. These alternative conceptions are more popular at lower ability levels than the scientifically correct answer. Sometimes the popularity of the alternative conception far exceeds that of the correct answer and the estimate of guessing, making them the most popular response at certain ability levels. For students at certain levels, alternative conceptions are not as popular, but have a substantial appeal, well above chance. The spread of "don't know," alternative conception, and correct answer in ability can be quite large.

Progress toward understanding key scientific concepts, such as the ones examined in the Project STAR Inventory, is not simple or straightforward. Students do not move quickly from

no opinion to the scientific understanding, but they do change. This movement is often painstakingly slow, taking much longer than a single year of study, and is most often mediated by belief in alternative conceptions. For example, most students moving from 8th to 12th grade would come to understand how the sun moves through the sky, but never fully connect it to and master the reason for seasons.

The curriculum represented in our nation's textbooks appears to be severely out of line with the results of this study. Concepts are often inappropriate for the grade level at which they are aimed, and mismatched to what students know and can learn. Many concepts are too difficult for students even at much higher grade levels. For example, the inverse square law of light is a concept that appears to reach a level of mastery only well after high school, yet it is included in the ninth-grade learning sequence for the NSTA's Scope Sequence and Coordination (p. 106). It is doubtful that most high school freshmen will succeed where Harvard undergraduates do not.

Rather than pursue the learning of such empirical laws and models directly, students may be better served by a more indirect approach. Since mastery of these concepts appears to take place only after many years, having students develop and test their own conceptions may have merit. Since so many high school students believe in an inverse first-power law for light, it would probably be more instructive for them to test their belief using a light meter or null photometer (Coyle, Gregory, Luzader, Sadler, & Shapiro, 1992). If they discover that their predictions are not reflected by their experiments, this can provide motivation for reformulating their ideas.

Fundamental concepts appear to take an extraordinarily long time to learn. Movement toward spiral curricula in science is encouraging, because it acknowledges research that shows systematically spaced teaching efforts have a significant and large effect on learning (Demster, 1992). Curricula must be constructed to teach and aid students in applying knowledge in different ways without appearing overly repetitive. When viewed as repeating something they have previously covered, students can lose motivation and interest. For example, learning the textbook explanation for moon phases year after year will most certainly be boring. Observing the moon through inexpensive telescopes, recording its position and appearance over monthly cycles, predicting its appearance and motions, learning the names of its surface features, calculating the heights of its mountains from their shadows, and estimating its size and distance from eclipse photographs can spread learning out over many years. This also gives students firsthand familiarity with phenomena, recordkeeping, and testing their ideas, all of which can only help the process of conceptual change. Such experiences give students much more evidence from which to make inferences and make them less dependent on authorities for their knowledge. This is certainly a positive step toward understanding the scientific process and how scientific knowledge is created.

The prior analysis helps to suggest a more reasonable ordering of concepts in astronomy. It makes sense to order from easy to difficult, and to deal with ideas that students have already spent time thinking about and developing on their own, even though they may not be scientifically correct. The structure revealed by this IRT analysis represents the natural flowering of students' ideas. Students create with a view that light "tires" before they imagine that it falls off in strength inversely to distance. It would make sense to discuss and experiment with these ideas as they occur, rather than impose totally foreign models which ignore student views. The analysis suggests several topics which may be far more appropriate than current middle school curricula for use prior to the eighth-grade level (< -0.1). These would be:

- what goes around what in the solar system (the dance of the planets),
- what moves in the sky (everything except the patterns of stars),

- how objects in the sky differ (changing shapes, orientations),
- how things appear different from different perspectives (spatial relations),
- experiments in the dark (at the bottom of foil-capped tubes),
- creating and playing with shadows (predicting shadow size and placement),
- making and reading each other's graphs (surveys and other data), and
- measuring angles and angular size (on earth and in the sky).

The analysis above also reveals a curious finding, that what we often think of as facts can actually be concepts. Teachers often speak of facts and the need for students to memorize them. For example, the star nearest the sun is 4 light-years away, modeled as two grapes at a hundred miles. Yet, the process of learning the distance between the sun and a neighboring star does not simply move from "don't know" to 4 light years. The correct answer does not appear to be fully formed. There is a marked preference for increasingly larger distances (as measured by the ratio of solar diameter to distance) as students age. The model including neighboring stars grows larger and larger in learning astronomy. Many such facts, especially quantitative ones, are connected to mental models. Learning such truths without paying attention to the models that they describe can result in conundrums such as Saturn being closer to us than the sun, or galaxies being closer than the stars. Such facts should not be taught in isolation, but only to help move students to more powerful and accurate models. Facts must fit into student's frameworks.

Pedagogical Issues

How should the structure of student knowledge affect teaching? For one, the teaching of key scientific concepts poses great difficulties. Student progress will be tortuous and at times often appear to regress. The J-shaped item response curves for the correct answers imply that students may temporarily perform less well by certain measures as they make progress toward ultimate understanding. As a result, test makers have avoided such questions, viewing them as problematic rather than seeing them as measuring real cognitive processes (Hutchinson, 1991). Teaching about seasons, day and night, or the nature of astronomical distances may appear to be detrimental to student understanding; youngsters who originally could repeat the generally accepted view will find alternative conceptions more appealing after instruction. Other students will very often strengthen their belief in alternative conceptions as they selectively reinforce their ideas from readings, lectures, labs, or demonstrations. It is important to keep in mind that this is not bad. It simply marks an initial step in meaningful learning; students often learn by first constructing the most plausible explanation for a phenomenon and then later find it wanting in predictive power. Students may move from one alternative conception to another, appearing to delay progress toward scientific understanding. This, again, is not bad. Moving from thinking that the seasons are caused by an elliptical orbit to that of a hemispheric distance differential is a positive step. Moving rightward on any of the graphs marks progress, whether it encompasses a scientifically accepted concept or not.

Student ideas must be honored in the classroom, for students must commit to their views before they can test them. This contrasts with teaching true facts, so-called rote learning, in which information is not integrated into students' conceptual frameworks. Such learning is cumulative (Novak & Gowin, 1984). Teachers should expect no regression if there are no conflicting student conceptions. This is often reflected in the tests constructed by many teachers who tend to prefer to test for rote learning rather than meaningful learning (Madaus, West, Harmon, & Lomax, 1992).

Abandonment of alternative conceptions without having a fully formed scientific conception can be quite frustrating for students and emotionally stressful. This period of confusion is

quite normal for students, especially in science, as they are attempting to break down and reconstruct their ideas (Lipson, 1993). Those students who have little experience in putting their own ideas to the test and, as a result, having to construct new ones will find such modes of learning quite distressing. It will often be the best students, those who have previously excelled in "quadrant two" learning (McCarthy, 1992; appeal of authoritative sources and passive acceptance of abstract ideas) who are the most vocal in their objections.

Although it is generally considered beneficial for teachers to have high expectations for their students, having expectations that are well beyond any realistically achievable progress for students can be a problem. It should be no surprise that many scientific concepts are difficult, since they represent the cumulative product of the brightest minds of the last several hundred centuries. Students find it quite trying to make such ideas their own. When students have not made meaningful changes to their knowledge structure, teachers often respond by restricting the range of assessment questions so that they more closely replicate the context and conditions of the classroom, limiting the applicability of the scientific knowledge. An overabundance of low-level questions will maintain high average scores on tests without assessing real understanding. Meaningful learning enables one to cross domains and apply ideas in previously unencountered contexts (Nisbett, Fong, Lehman, & Cheng, 1987). Distractor-driven multiple choice questions which pit students' previous ideas versus scientific views are very effective in examining conceptual change. Other, more open-ended assessment tools that incorporate alternative conceptions (such as an essay question that asks for an explanation of why the earth's elliptical orbit fails to explain the seasons, or interpreting the opinions of siblings concerning such topics), are less restrictive and easier to construct.

Teachers who attempt to change students' ideas will find that coverage will suffer. It is not possible to include all of the concepts in a typical precollege textbook in a yearlong precollege science course. Teachers will find that they must spend more time on concepts and must revisit ideas covered in previous years. Limiting coverage does not appear to be detrimental; the author's recent work on the impact of high school preparation on college physics success shows that conventional soup-to-nuts high school physics courses appear to have little to do with college physics success, whereas courses that cover few topics in great depth serve as much more effective preparation (Sadler, 1994).

Limitations of the Study

This study represents an alternative from the primarily qualitative analysis of children's ideas and the quantitative studies that rely on classical test theory for item analysis. Our analyses require large datasets to estimate all parameters (three for each answer and one plus the number of answers per item. In this case, five answers plus the DK category times three, making 18 per item). The minimum number of subjects for a three-parameter logistic model with a latent "don't know" category is given by this simple expression along with the calculation for the Project STAR test:

$$N > \text{items} * (3 * \text{answers/item} + 2) = 3 * \text{answers} + 2 * \text{items}$$
$$N > 3 * 47 * 5 + 2 * 47 = 799 \quad (2)$$

The population must not only be representative of level of interest, but must also include subjects higher and lower than the intended range so that the functions will reach their asymptotes. The measure of statistical significance that applies to the goodness of fit is the standard error in measurement of the estimated ability.

The conclusions of this study depend upon the assumption that cross-sectional data can support a longitudinal interpretation (Millar et al., 1993). While this may seem reasonable, only longitudinal studies can provide validation for the transitions that individuals experience. The use of such longitudinal tests is complicated by the fact that the measured ability change is small (0.13 SD without instruction, 0.16 with traditional instruction, and 0.26 with Project STAR), especially in comparison with the average variation in each grade (0.73 SD). If the population changes very slowly, as is the case with most conceptual learning, pre- and posttesting does not reveal much change in student understanding. The variation in students' performance within a grade is much larger than between grades. In this case, a study of the population as a whole may allow inferences to be made about long term growth which would take a great many years (about 5 in this population) to detect through a longitudinal study. Adding in expert subjects (college professors, graduate students) would make longitudinal studies of this type prohibitive in their duration and cost.

A hypothetical example of such a cross-sectional study would be to predict the mortality of heretofore undiscovered trees in a tropical rainforest by visiting and measuring their health yearly. After 300 years, the dataset would show the fraction that survive to old age. However, if one were to examine the data of a single year of such old-growth forest, one could construct a model of mortality based on the existing population of old trees and younger ones. The method is fraught with problems (mainly assuming conditions were always as they are now), but it gives a glimpse of order that is difficult to find any other way and may aid further study.

The only way we can tell if a student has some understanding is by using an indirect measure. We must observe the student's written or spoken responses, dialogue with peers, or performance in laboratory or real-life situations (Millar et al., 1993). In the domain of students' scientific conceptions, interviews have provided the largest and clearest window. Multiple choice tests based on interview data may do less well at assessing student understanding, particularly if test writers misinterpret qualitative studies or such studies are inadequate.

Opportunities for Future Research

Efforts at curriculum reform are widespread at the national, state, and city levels. Many assessment strategies have been proposed that draw on innovations in alternative assessment, such as portfolios, journals, and projects. The inclusion of distractor-driven multiple choice tests provides a unique opportunity to have a highly reliable and comparable tool. The development of such a test would aid teachers in diagnostic work, provide formative feedback to curriculum designers, and allow summative evaluations of curricular experiments.

Conceptual change appears to take a long time. A spiral curriculum implicitly recognizes this fact that it takes many years to master key concepts. An IRT cross-sectional analysis is more sensitive to small changes in learning than pre- and posttesting over a short time period because it spreads students out over a vast continuum, one that is far larger than a year's change. It would be useful to compare this cross-sectional analysis with a carefully controlled longitudinal study to find its strengths and its weaknesses.

This study examined student ideas in astronomy. Although some ideas are shared with physics, it is not known whether similar behaviors are manifest in other scientific disciplines. It would be useful to carry out similar procedures on popular multiple choice alternative conceptions tests in astronomy (Lightman & Miller, 1989) and in other domains such as the force concept inventory (Hestenes et al., 1992).

The items on the Project STAR test are plotted along a single scale. Items may also exhibit additional structure by being prerequisites for each other. For example, correctly answering

Item 1 might be a necessary but not sufficient condition for correctly answering for Item 17. Selection of distractors may be related to the selection of other distractors. The possibility for this kind of analysis has been explored previously, building a knowledge structure from the relationship between questions and alternative conceptions (Bart & Krus, 1973; Haertel & Wiley, 1993; Sadler, 1995). In addition, other statistical methods such as latent class analysis may be useful in combining similarly functioning items into groups in a purely objective fashion.

Conclusions

Multiple choice tests coupled with psychometric tools can become powerful windows into children's ideas in science. Useful multiple choice tests can be constructed from qualitative investigations of students' conceptions in science. Such tests are reliable and valid tools for assessing the relative popularity of alternative ideas. By applying the methods of item response theory, both scientifically correct answers and alternative conceptions can be modeled.

Correct answers to such questions have very unusual psychometric profiles. The probability of choosing the scientific answer surprisingly decreases with student ability before it rises to unity with high-level students. Such behavior has been noted in the literature as an oddity, as a possible sign of cheating or bungled construction. Never has an entire test revealed signs of such nonmonotonicity. In this instrument, one that attempts to test for student alternative conceptions, it appears that students at moderate levels of ability actually retreat from the scientific explanation. Analysis of the incorrect responses, which are drawn from the alternative conception literature, shows this quite clearly. Students of moderate ability often find alternative conceptions increasingly attractive. Evidence for alternative conceptions being intermediate and positive steps in the processes of coming to understand scientific concepts is shown by plotting items and distractor difficulties on a single uniform scale of ability based upon overall test score.

The difference between the ability level of students who are guessing at answers and those who have mastered the concept can be thought of as a parameter that is related to the length of time it takes to learn the concept. In addition, entire populations of students can be characterized by their position on the ability continuum. IRT allows widely disparate groups to be compared. Students in Grades 8–12, college astronomy majors, graduate students, and professors have been included in this analysis. The results of pre- and posttests from control and experimental classes have also been contrasted.

By examining the grade-level responses of students and the impact of an introductory-level course in astronomy on their responses, the time period for conceptual change turns out to be much longer than previously thought. A single year is a very short time to reconstruct students' ideas; most concepts appear to take years. The difficulty of many concepts is so high that it is likely most students who are exposed to an idea once have virtually no chance of mastering it. An example is the cause of seasons, which is covered in Grades 5 or 6, and which only reaches a level of mastery in college astronomy majors. By comparing teacher predictions of student scores on the same scale, one finds that teachers are quite accurate in their estimates of students' starting knowledge but optimistic in the extreme about the impact of their own teaching. Students appear to gain only one eighth of what teachers predict for their courses.

Efforts to revise or create new curricula (such as the current national efforts) do not appear to take the extreme difficulty of scientific ideas into account. Many of the astronomical ideas found in textbooks and the new standards are far too difficult; students do not appear to have learned them in the past. Curriculum and pedagogical changes in the future offer no guarantee that students will be able to master these ideas while in high school. However, some ideas do change for some students during high school. These are concepts which could be appropriate

for middle school students' study. These concepts may be learned more quickly or easily than they are at present. One positive step is that the new standards revisit key ideas frequently, and this spiral approach may more closely match the natural timetable in which students' ideas progress.

Modeling the changing ideas of large groups of students begins to tie cognitive theory to student performance in a quantitative and significant fashion. There is evidence that students' alternative conception appear to get stronger before they recede and that such sequences of understanding are stable and predictable. Moreover, since cognitive change is so slow, it may be useful to examine alternative pedagogies such as explicitly dealing with alternative conceptions by attempting to shore them up and strengthen them. Teachers could then help to dismantle them later with key evidence. If certain alternative conceptions are intermediate steps in the learning of key scientific ideas, one must consider whether teaching them explicitly may be of considerable utility.[12]

For the future, longitudinal studies that explore in detail the patterns and models of cognitive change presented in this two-time cross-sectional study would be very useful. Studies of other multiple choice tests of students' ideas in other disciplines are currently under way by the author. These can serve as comparisons that will further test the application of psychometric methods to cognitive change. Such work will also test the degree to which the sequence and structure of learning in astronomy is similar to other sciences.

Based on the evidence provided by this study, several steps should be considered by the science education community:

- Distractor-based multiple choice tests should be created to aid teachers in diagnosing student conceptions and to easily measure conceptual change.
- Standardized tests must be revised or created anew, incorporating distractors which fairly reflect the results of qualitative studies.
- Nonmonotonic psychometric models should be used to assess tests of scientific understanding. Scoring models should reflect students' stagelike progression in conceptual understanding.
- Test results should be used by standards developers and curriculum developers to provide a baseline for understanding of science keyed to grade level.
- New curricula should discuss and treat alternative conceptions not as errors, but as stepping stones to scientific understanding.

Qualitative studies have provided the foundation for the development of reliable and powerful distractor-driven multiple choice assessments of students ideas in several domains. Interviews have documented the major alternative conceptions held by students. Tests built from these studies are characterized by their very attractive distractors. Such answers are characteristically removed from items by standardized test makers because they do not fit the accepted psychometric profiles. This reduces the utility of most written tests to measure conceptual understanding. Alternative profiles that reflect what we know of cognitive development are essential if tests are to reflect meaningful learning and not only rote memorization. Easy-to-score tests of conceptual understanding are essential elements in the reform of science teaching. It is essential that such tests be developed and that standardized tests now being created or revised use the bounty of our rich legacy of qualitative research on children's ideas in science.

This research was supported by grants from the National Science Foundation (MDR 85-50297 and MDR 88-50424), the Smithsonian Institution, and Apple Computer. The authors thanks his Harvard and Smithsonian colleagues, Brian Alters, Hal Coyle, Bruce Gregory, Roy Gould, Bill Luzader, Judith Peritz,

Susan Roudebush, Matt Schneps, Irwin Shapiro, Judy Singer, Robert Tai, and Terrence Tivnan, for their comments and support. Marcus Lieberman helped in formulating this research agenda and carried out the multiple IRT calculations with great patience and good humor. Project STAR teachers from across the United States have contributed immensely by agreeing to open their classrooms to this research. Thanks also go to Joel Mintzes at University of North Carolina, Wilmington; Joe Novak at Cornell; Yossi Nussbam at Michlalah Jerusalem College for Women; David Hamer at Tufts University; and Ken Schoon at Indiana University Northwest, for their feedback and encouragement.

Notes

[1] The Private Universe Teleconference was broadcast as nine 2-hour programs in the fall of 1994. It was produced by the author's colleague, Matt Schneps, of the Harvard-Smithsonian Center for Astrophysics and sponsored by Annenberg/Corporation for Public Broadcasting and the National Science Foundation.

[2] "If I am interested in individuals' understandings, I see how they talk and perform activities. I don't con them into providing answers that someone later uses to establish expert–novice dichotomies" (anonymous reviewer's comment condemning an earlier paper submitted to the *International Journal of Science Education* in 1995).

[3] Now published by Kendall/Hunt, Dubuque, IA.

[4] The most well-known of these interviews were made into the award-winning video, *A Private Universe* (Schneps and Sadler, 1988), which captures the ideas of graduating Harvard seniors and younger students about seasons and the phases of the moon. It is currently distributed by Annenberg/CPB and by the Astronomical Society of the Pacific.

[5] Special thanks go to teaching colleagues Jennifer Bond and Paul Hickman of Boston University Academy and Northeastern University, respectively, and to Anne Young of Rochester Institute of Technology, for their work on light and color interviews.

[6] A cross-age study of students' ideas of electrical circuits found the "sequence model" to be very popular [Shipstone, D. (1985). Electricity in simple circuits. In R. Driver, E. Guesne, & A. Tiberhien (Eds.), *Children's ideas in science* (pp. 33–51). Philadelphia: Open University Press]. This model treats electricity flow through electrical components in a circuit as if electricity is consumed by components in sequence. This model had the same popularity (35%) in both 12- and 17-year-old students. However, the popularity of the model climbed to 80% in 14-year-olds before it ultimately receded.

[7] Written by David Thissen, the program is published by Scientific Software, at 1525 E. 53rd Street, Suite 906, Chicago, IL 60615.

[8] An item similar to the one constructed for this test was included in the 1969 National Assessment of Educational Progress of 9-year-old (third-grade) students. The percentage choosing each answer follows in parentheses: One reason that there is day and night on earth is that the:

Sun turns. (8%) Moon turns. (4%) Earth turns. (81%)

Sun gets dark at night. (6%) I don't know. (1%)

This is a fine example of a question that avoids misconceptions. The high percentage of correct answers can be attributed to the lack of plausible distractors (Schoon, 1988).

[9] E.g., let us consider students who are assigned an ability of -0.5 from answering other items and who choose the hemispheric differential answer to Item 17.

[10] This could be done through building scale models of the earth–sun system and finding that their distance changes by only $\pm 1\%$ over a year. A light meter shows virtually no change in measured intensity on this scale, using a lamp as the sun.

[11] A clear 200-W lamp for the sun, students' heads for the earth, and a tennis ball as the moon works well. With the ball roughly 20 ft from students, one can demonstrate that during its monthly excursion about the earth, the moon can only be in the earth's shadow for less than a day (a lunar eclipse) and be partially blocked for a very short time. Since the moon's shape is pretty much constant over a night and cycles over a month, the earth's shadow cannot have a role in phases.

[12] This can be done tongue in cheek, as I have pretended to be a second-grade student with typical views. At workshops, teachers have tried to convince me that the earth is a sphere, not flat. After reject-

ing authority repeatedly, participants take most of an hour to reconstruct, from their own stock of evidence, the ancient Greek arguments for a spherical earth.

References

Aldridge, B.G. (1995). *A high school framework for national science education standards.* Arlington, VA: National Science Teachers Association.

Baker, F.B. (1985). *The basics of Item Response Theory.* Portsmouth, NH: Heineman.

Bart, W.M., & Krus, D.J. (1973). An ordering-theoretic method to determine hierarchies among items. *Educational and Psychological Measurement, 33,* 291–300.

Baxter, J. (1989). Children's understanding of familiar astronomical events. *International Journal of Science Education, 11,* 502–513.

Bell, A., Brekke, G., & Swan, M. (1987). Alternative conceptions, conflict and discussion in the teaching of graphical interpretation. In J.D. Novak (Ed.), *2nd International Seminar on Misconception and Educational Strategies in Science and Mathematics* (Vol. 1) (pp. 46–58). Ithaca, NY: Cornell University Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bjar, I.I.N. (1993). A generative approach to psychological and educational measurement. In N. Fredericksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–357). Hillsdale, NJ: Erlbaum.

Bock, R.D. (1972). Estimating item parameters and latent proficiency when the responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Borko, H., & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal, 26,* 473–498.

BouJaoude, S.B. (1992). The relationship between students' learning strategies and the change in their misunderstandings during a high school chemistry course. *Journal of Research in Science Teaching, 29,* 687–699.

Bouwens, R.E.A. (1986). Alternative conceptions among pupils regarding geometrical optics. In J. Hunt (Ed.), *GIREP Conference: Cosmos—an educational challenge* (pp. 369–370). Copenhagen: European Space Agency.

Choppin, B. (1985). A two-parameter latent trait model. In H.J. Walberg & T.N. Postlethwaite (Eds.), *Evaluation in education* (pp. 43–62).

Council, N.R. (1996). *National Science Education Standards.* Washington, DC: National Academy Press.

Coyle, H., Gregory, B., Luzader, W., Sadler, P., & Shapiro, I. (1993). *Project STAR: The universe in your hands.* Dubuque: Kendall/Hunt.

Demster, F.N. (1992). The spacing effect: A case study in the failure to apply the results of psychological research. In M.K. Pearsall (ed.), *Scope sequence and coordination of secondary school science* (pp. 25–38). Washington, DC: National Science Teachers Association.

diSessa, A. (1982). Unlearning Aristotelian physics: A study of knowledge-based learning. *Cognitive Science, 6,* 37–75.

diSessa, A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10,* 105–225.

Doig, B., & Adams, R. (1993). Methodological alternative conceptions: Naive or not interested? In J. Novak (Ed.), *Proceedings of the third international seminar on Alternative Con-*

ceptions and Educational Strategies in Science and Mathematics. Ithaca, NY: Alternative Conceptions Trust.

Driver, R. (1973). *The representation of conceptual frameworks in young adolescent science students.* Ph.D. dissertation, University of Illinois.

Duckworth, E. (1987). "*The having of wonderful ideas*" *and other essays on teaching and learning.* New York: Teachers College Press.

Dufresne, R., Gerace, W., Hardiman, P.T., & Mestre, J. (1986). Hierarchically structured problem solving in elementary mechanics: Guiding novices' problem analysis. In J. Hunt (Ed.), *GIREP Conference: Cosmos—an educational challenge* (pp. 116–130). Copenhagen: European Space Agency.

Eaton, J.F. (1984). Student alternative conceptions interface with science learning: Case studies of fifth-grade students. *Elementary School Journal, 84,* 365–379.

Eckstein, S.G., & Shemesh, M. (1993). Stage theory of the development of alternative conceptions. *Journal of Research in Science Teaching, 30,* 45–64.

Freyberg, P., & Osborne, R. (1985). Constructing a survey of alternative views. In R.J. Osborne & P. Freyberg (Eds.), *Learning in science: The implication of childrens' science* (pp. 166–167). Auckland, New Zealand: Heineman.

Furuness, L.B., and Cohen, M. (1989, April 1). *Children's conception of the seasons: A comparison of three interview techniques.* Paper presented at the National Association for Research in Science Teaching Meeting.

Haertel, E.H., & Wiley, D.E. (1993). Representations of ability structures: Implications for testing. In N. Fredericksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359–384). Hillsdale, NJ: Erlbaum.

Haladyna, T.M. (1994). *Developing and validating multiple-choice test items.* Hillsdale, NJ: Erlbaum.

Halloun, I.A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics, 53,* 1043–1055.

Hambleton, R.K. (1989). Principles and selected applications of Item Response Theory. In R.L. Linn (Eds.), *Educational measurement* (pp. 147–200). New York: Macmillian.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage.

Hammer, D. (1995). *Alternative conceptions or P-prims: How might alternative perspectives of cognitive structure influence instructional perceptions and intentions?* Newton, MA: Center for the Development of Teaching, Education Development Center.

Hardiman, P.T., Dufresne, R., & Gerace, W. (1986). Physics novices' judgments of solution similarity: When are they based on principles? In J. Hunt (Ed.), *GIREP conference: Cosmos—an educational challenge* (pp. 194–202). Copenhagen: European Space Agency.

Hestenes, D., Wells, M., & Swackhammer, G. (1992). Force concept inventory. *The Physics Teacher, 30,* 141–158.

Hutchinson, T.P. (1991). *Ability, partial information, guessing: Statistical modeling applied to multiple-choice tests.* Adelaide, Australia: Rumsby Scientific.

Jung, W. (1987). Understanding students' understandings: The case of elementary optics. In J.D. Novak (Ed.), *2nd international seminar on Misconception and Educational Strategies in Science and Mathematics, 3* (pp. 268–277). Ithaca, NY: Cornell University Press.

Klein, C. (1982). Children's concepts of the earth and sun: A cross cultural study. *Science Education, 65,* 95–107.

Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Lightman, A., & Sadler, P.M. (1993). Teacher predictions versus actual student gains. *The Physics Teacher, 31,* 162–167.

Lightman, A.P., & Miller, J.D. (1989). Contemporary cosmological beliefs. *Social Studies of Science, 19,* 127–136.

Linn, M., (1986). Science. In R.F. Dillon & R.J. Sterberg (Eds.), *Cognition and instruction* (pp. 155–204). New York: Academic.

Lipson, A. (1993). The confused student. *College Teaching, 40,* 91–95.

Loria, A., Michelini, M., and Mascellani, V. (1986). Teaching astronomy to pupils aged 11–13. In J. Hunt (ed.), *GIREP Conference: Cosmos: an educational challenge* (pp. 229–233). Copenhagen: European Space Agency.

Madaus, G.F., West, M.M., Harmon, M.E., & Lomax, R. (1992). *The influence of testing on teaching math and science in Grades 4–12.* No. SPA 8954759). Boston: Center for the Study of Testing and Evaluation, Boston College.

Marshall, K., & Lancaster, O. (1983). *Science: Elementary and middle school curriculum objectives.* Boston: Boston Public Schools.

Masters, G.N. (1988). An analysis of partial credit scoring. *Applied Measurement in Education, 1,* 279–297.

Masters, G.N., & Mislevy, R.J. (1993). New views of student learning: Implications for educational measurement. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 219–242). Hillsdale, NJ: Erlbaum.

McCarthy, B. (1992). *4 MAT.* Excel.

Messnick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13–104). New York: American Council on Education.

Michaels, S. (1995). *Narratives and inscriptions: Cultural tools, power, and powerful sensemaking.* Cambridge: National Academy of Education, Symposium on Anthropologists' Perspectives on Learning Science in and out of School.

Millar, R., Gott, R., Lubben, F., & Duggan, S. (1993). *Children's performance of investigative tasks in science: A framework for considering progression.* Liverpool: British Educational Research Association. (ERIC Document Reproduction Service No. ED364417)

Mislevy, R.J. (1994). *Test theory reconceived: Project 2.4 quantitative models to monitor the status and progress of learning and performance and their antecedents.* National Center for Research on Evaluation, Standards and Student Testing. (Eric Document Reproduction Service No. ED376180)

Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrica, 55,* 195–215.

Narode, R. (1987). Standardized testing for alternative conceptions in basic mathematics. In J.D. Novak (Ed.), *2nd international seminar on Misconception and Educational Strategies in Science and Mathematics* (Vol. 1) (pp. 222–333). Ithaca, NY: Cornell University Press.

Newman, D., & Morrison, D. (1993). The conflict between teaching and scientific sense-making: The case of a curriculum on seasonal change. *Interactive Learning Environments, 3,* 1–16.

Nisbett, R.E., Fong, G.T., Lehman, D.R., & Cheng, P.W. (1987). Teaching reasoning. *Science, 238,* 625–631.

Novak, J.D., & Gowin, D.B. (1984). *Learning how to learn.* Cambridge: Cambridge University Press.

Nussbaum, J. (1979). Childrens' conception of the Earth as a cosmic body: A cross age study. *Science Education, 63,* 83–93.

Nussbaum, J. (1985). The earth as a cosmic body. In R. Driver, E. Guesne, & A. Tiberhien (Eds.), *Children's ideas in science* (pp. 170–192). Philadelphia: Open University Press.

Nussbaum, J., & Novak, J. (1982). Alternative frameworks, conceptual conflict and accommodation: Toward a principled teaching strategy. *Instructional Science, 11,* 183–200.

Osborne, R.J., & Gilbert, J.K. (1980). A technique for exploring students' views of the world. *Physics Education, 15,* 376–379.

Pfundt, H., & Duit, R. (1985). *Bibliography: Students' alternative frameworks and science education.* Kiel, UK: IPN.

Project 2061, American Association for the Advancement of Science. (1993). *Benchmarks for Science Literacy.* New York: Oxford University Press.

Reif, F. (1987). Instructional design, cognition, and technology: Applications to the teaching of scientific concepts. *Journal of Research in Science Teaching, 24,* 309–324.

Richards, D.D., & Siegler, R.S. (1982). U-shaped behavioral curves: It's not whether you're right or wrong, it's why. In S. Strauss (Eds.), *U-shaped behavioral growth* (pp. 37–61). New York: Academic Press.

Sadler, P.M. (1987). Alternative conceptions in astronomy. In J.D. Novak (Ed.), *2nd international seminar on Misconception and Educational Strategies in Science and Mathematics* (Vol. 3) (pp. 422–425). Ithaca, NY: Cornell University Press.

Sadler, P.M. (1992). *The initial knowledge state of high school astronomy students.* Ed.D., Harvard Graduate School of Education, Cambridge.

Sadler, P.M. (1994, 16 January). *The effect of high school preparation on college physics success.* Paper presented at American Association of Physics Teachers Annual Conference, Orlando.

Sadler, P.M. (1995). Astronomy's conceptual hierarchy. In J. Percy (Ed.), *Proceedings of the ASP Astronomy Education Meeting.* San Francisco: Astronomical Society of the Pacific.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Psychometric Society.

Schneps, M.H., & Sadler, P.M. (1988). *A private universe.* Santa Monica, CA: Pyramid Films.

Schoon, K.J. (1988). *Alternative conceptions in earth and space sciences: A cross-age study.* Ph.D. dissertation, Loyola University.

Sherman, S.W. (1976, April). Multiple-choice test bias uncovered by the use of an "I don't know" alternative. In *annual meeting of the American Educational Research Association,* Chicago.

Shipstone, D. (1985). Electricity in simple circuits. In R. Driver, E. Guesne, & A. Tiberhien (Eds.), *Children's ideas in science* (pp. 33–51). Philadelphia: Open University Press.

Shipstone, D.M., Rhoneck, C., Jung, W., Karrqvist, C., Dupin, J.J., Joshua, S., & Licht, P. (1987). A study of students' understanding of electricity in five European countries. *European Journal of Science Education.*

Smith, J., DiSessa, A., & Rochelle, J. (1993). Alternative conceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences, 3,* 115–163.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49,* 501–519.

Thissen, D., Steinberg, L., & Fitzpatrick, A.R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26,* 161–176.

Touger, J.S. (1985, 23 January). *Students' conceptions about planetary motion.* Paper presented at the American Association of Physics Teachers Meeting.

Vosniadou, S., & Brewer, W.F. (1987). Theories of knowledge restructuring in development. *Review of Educational Research, 57,* 51–67.

Wandersee, J.H. (1986). Can the history of science help science educators anticipate students' misconceptions? *Journal of Research in Science Teaching, 23,* 581–597.

Weiss, I.R. (1987). *Report of the 1985–86 National Survey of Science and Mathematics Education* (No. RTI/2938/00-FR). Cary, NC: Research Triangle Institute.

Welch, W.W., Harris, L.J., & Anderson, R.E. (1984). How many are enrolled in science? *The Science Teacher, 51*, 16.