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

Astronomy Education Review

Volume 1, Oct 2001 - Jan 2002
Issue 1

Development of the Astronomy Diagnostic Test

by **Beth Hufnagel**
Anne Arundel Community College
Posted: 06/14/02

The Astronomy Education Review, Issue 1, Volume 1:47-51, 2002

Abstract

This paper describes the development of the Astronomy Diagnostic Test (ADT), a 33-question pencil-and-paper multiple-choice survey, up to the release of Version 2.0 in June 1999. The 21 content questions of the ADT (1) address concepts included in most introductory astronomy courses for non-science majors, (2) include only concepts recognizable to most high school graduates, (3) focus on one concept only, and (4) avoid jargon. Statistical analysis based on hundreds of students? responses, 30 open- ended written responses, and student interviews guided the wording and final selection of these questions.

1. INTRODUCTION

In July 1998, a group of astronomy education researchers met in Albuquerque, New Mexico at a meeting of the Astronomical Society of the Pacific and discovered that the community of astronomy instructors needed a standard assessment instrument to measure the understanding of students taking introductory undergraduate astronomy classes for non-science majors. We formed the Collaboration for Astronomy Education Research (CAER) with the goal of producing a multiple-choice, education research-based assessment survey. The CAER included Timothy Slater, Grace Deming, Jeff Adams, Rebecca Lindell (Adrian), Christine Brick, Michael Zeilik, the author, and a number of other experts who helped as needed.

The theory underlying our research was that student learning during an introductory astronomy course is closely linked to prior knowledge and beliefs about the universe around them, i.e., their understanding of the content of astronomy. To enhance learning, as Hammer (1996) puts it, instructors must "attend to what their students are doing." (Other aspects of student knowledge, e.g., the nature of science, what does "knowing" mean, how the students view themselves as learners and scientists, and how others will view them as scientists (the nerd effect), were not the main focus of this instrument.) This theory guided the development of a pencil-and-paper multiple-choice assessment tool that could be used to measure the level of the students' understanding before they took the course, without discouraging or frightening them. This

means that the content had to be limited to what is often taught in elementary and high school classrooms in the United States. To make it attractive to instructors and professors not accustomed to standard tests, we also wanted an assessment tool that wouldn't take up much in-class time.

2. DEVELOPMENT OF THE ASTRONOMY DIAGNOSTIC TEST

2.1. Predecessor Instruments

There were two predecessors for this new diagnostic. The first was Phil Sadler's 47-item Project STAR Astronomy Concept Inventory (Sadler 1998). Because it was developed for high school students, the content was appropriate and at the right level. However, it was not widely available, and its validity and reliability had never been measured for undergraduates, our intended users. The second was Zeilik's initial 1998 version of the Astronomy Diagnostic Test (ADT Version 1.0; Zeilik 1998a), which consisted of 13 questions from his Misconceptions Measure (Zeilik et al. 1997), 10 questions from the Project STAR Astronomy Concept Inventory, and 10 new questions.

Zeilik's 1998 version of the ADT was re-written by the CAER using standard psychometric principles, e.g., Miyasaka and Ryan (1997). These principles for multiple-choice tests include having only one concept per question, enabling the correct answer to be known before reading the answers, and avoiding scientific jargon. An example of a question about stellar evolution from an early ADT version is presented in Figure 1, and one from the final ADT of June 1999 (Version 2.0) is given for contrast in Figure 2. In addition, the answers to Zeilik's 1998 ADT version had been published in Zeilik et al. (1998b) which, given the ability of the college students to search the Internet, may have impaired its usefulness. Additional questions were added from a previously unpublished survey by a member of the CAER group, Grace Deming, and from unpublished independent student interviews by Rebecca Lindell; the final topics included on the June 1999 ADT are listed in Figure 3.

Sample ADT Version 1.0 Question

Stars shine by nuclear fusion. Their lives end when their fusion reactions cease. More massive stars give off much more energy per second than less massive ones. What do you predict about the lifetimes of two stars, one more massive than the other?

A. They will both have the same lifetimes.
B. The more massive star will have a longer lifetime.
C. The less massive star will have a longer lifetime.

Figure 1. This question does not comply with standard psychometric principles because it includes more than one key idea, and jargon such as "nuclear fusion."

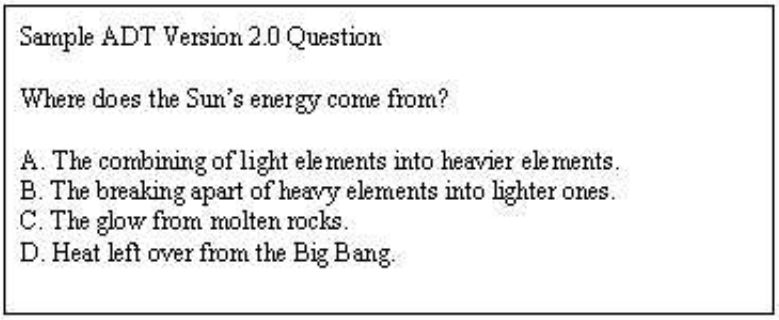

Figure 2. This question complies with psychometric standards because it includes only one idea, could be answered without looking at the suggested choices for the answers (distractors), and uses only phrases that students volunteered in their interviews. It is also a concept included in the National Science Standards (Adams & Slater 1999).

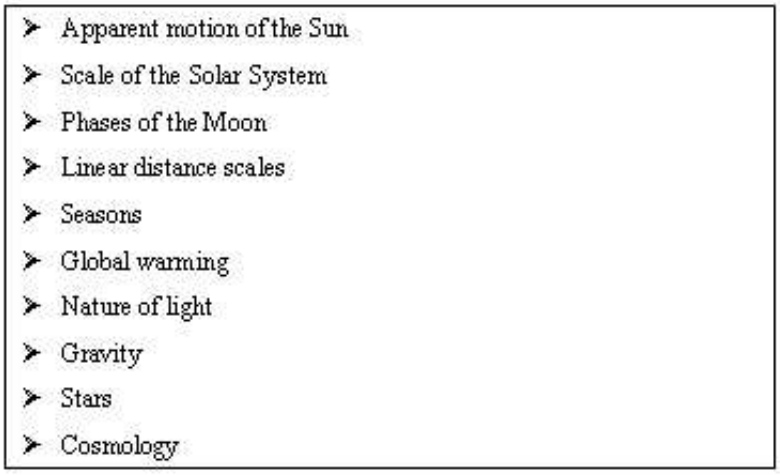

Figure 3. Topics on the ADT Version 2.0.

2.2. Reliability

The next task was to measure the reliability of the ADT (Creswell 1994). Two questions we asked were "Does a wrong answer mean that the student doesn't understand the concept being tested?" and "Does a correct answer mean that the student does understand the concept being tested?" Of course, the latter is more difficult to determine, as a student can guess the right answer. We took three complementary approaches: statistical analysis of hundreds of students taking the ADT in 34 classes across the continental U.S.; 30 written responses of students who were given the questions only; and about 60 interviews of University of Maryland and Montana State University students enrolled in ASTRO 101, the introductory astronomy course for non-science majors.

One example of the statistical analysis used was item discrimination; this identifies questions where students with otherwise high total average scores on the ADT consistently answer one question incorrectly, and vice versa. These items were removed from the next version of the ADT. We also applied more sophisticated techniques, such as Lei Bao's (2000) concentration-density (S-D) method to identify questions reflecting underlying common student models. Bao's analysis also highlighted the heterogeneity of the underlying population. For example, there are three or four questions where the statistic for men was acceptable, but the same statistic for these questions indicated that many women were guessing. We chose to retain such questions to measure the effectiveness of classroom reforms intended to close this gender gap. Data by institutional type, class size, and gender are in Hufnagel et al. (2000). Grace Deming's article, Results from the Astronomy Diagnostic Test National Project, appears in this issue of the *AER* and discusses additional statistical reliability studies conducted after the ADT version 2.0 was released.

The open-ended written and oral responses were used to identify items where the question was interpreted by the students in a way different than was intended, or the question tested a concept about which the students knew nothing. These student responses were also the sources for many of the incorrect responses adopted in the final version. One of the reasons the average score of the ADT is so low is that many students find what they believe to be the correct answer expressed in words familiar to them, i.e., their incorrect ideas are expressed in their own "natural language." Indeed, many students would score higher if they answered at random. This is also why the ADT is uniquely tailored to the American student. Many astronomy courses as currently presented do not change these initial student ideas as much as many professors would like; this is summed up by a quote from a university student in the last week of an astronomy course after I reviewed her answers to the ADT with her: "I seem to know, or think I know, a lot of things. I just don't know ? the insides of them."

2.3. Validity

Another issue is the validity of the ADT (Creswell 1994). Experienced astronomy professors familiar with the national standards for pre-college preparation developed it, but we also solicited and considered comments from other non-CAER participants who are experienced in teaching and astronomy education research. To be valid, the ADT should be able to measure the continuum of astronomy sophistication. This means that beginning students should score the lowest, advanced students should score higher, and experts (professors) should score close to 100%. This is true for the ADT, as during the time frame of development, the mean for undergraduates was 34%; the mean for a class of non-science majors taking a third astronomy course at the University of Maryland was 66%; and the mean was 97% for a group of professors. The only group of ASTRO 101 students to achieve similar scores to the advanced non-science majors consisted of men in a Midwest technical university. The women's average in this same class was statistically indistinguishable from that of other women.

3. SUMMARY

The end result of this effort by CAER was the ADT Version 2.0, released in June 1999. It has 33 multiple-choice questions, with 21 questions probing the astronomy concepts or knowledge held by students, and another 12 questions recording their background or attitudes. All of the questions and answers are phrased in the students' own natural language and have acceptable validity and reliability as defined by the educational community. The wrong answers, or distractors, reflect ideas held by many of the U.S.'s undergraduates taking astronomy courses for non-science majors. This work was supported in

part by National Science Foundation grants DGE-0003022 (BH), DGE-9714489 (BH), and REC-0089239 (GD). I am also grateful to the professors and instructors who participated in the development of the ADT, and for the cooperation of their students.

References

Bao, L. 1999, Dynamics of Student Modeling: A Theory, Algorithms, and Application to Quantum Mechanics, Ph.D. thesis, University of Maryland, 31.

Creswell, J. W. 1994, *Research Design Qualitative & Quantitative Approaches*, Thousand Oaks, Calif.: Sage Publications, 121.

Hammer, D. 1996, More than Misconceptions: Multiple Perspectives on Student Knowledge and Reasoning, and an Appropriate Role for Education Research, *American Journal of Physics*, 64(10), 1316.

Hufnagel, B., Slater, T., Deming, G., Adams, J., Lindell Adrian, R., Brick, C., & Zeilik, M. 2000, Pre-course Results from the Astronomy Diagnostic Test, *Publications of the Astronomical Society of Australia*, 17:2.

Miyasaka, J. R., & Ryan, J. M. 1997, Improving Student Assessment Strategies, *Big Sky Institute Professional Development Workshop Series*, September 11-12.

Sadler, P. M. 1998, Psychometric Models of Student Conceptions in Science: Reconciling Qualitative Studies and Distractor-driven Assessment Instruments, *Journal of Research in Science Teaching*, 35(10), 265.

Zeilik, M. 1998, private communication.

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

Zeilik, M., Schau, C., Mattern, N., Hall, S., Teague, K. W., & Bisard, W. 1997, Conceptual Astronomy: A Novel Model for Teaching Postsecondary Science Courses, *American Journal of Physics*, 65(10), 987.

ÆR
47 - 51