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

Astronomy Education Review

Volume 1
Issue 2

Birth of the Astronomy Diagnostic Test: Prototest Evolution

by **Michael Zeilik**
University of New Mexico

The Astronomy Education Review, Issue 2, Volume 1, 0

Abstract

The Astronomy Diagnostic Test (ADT) version 2 is the first research-based conceptual assessment developed for use in undergraduate introductory astronomy classrooms. Here I present the background of the development of earlier versions of the ADT as a basis for understanding the effort leading up to the ADT 2.

The following was presented as part of the Special Session, "The Astronomy Diagnostic Test: Development, Results, and Applications" at the January 2002 Meeting of the American Astronomical Society.

1. INTRODUCTION

In 1992, I received a grant from the National Science Foundation (DUE 92-53983) to reform our introductory astronomy course for non-science majors, "Astro 101." The goal was to transform the course from a mainly descriptive one to a primarily conceptual one, using results from physics and astronomy educational research and from the cognitive sciences (Bisard & Zeilik 1998). I assembled an interdisciplinary team: Candace Schau and Nancy Mattern (Educational Foundations, College of Education); Kathleen ("Kim") Teague (Cognitive Sciences, Dept. of Psychology), and Shannon Hall and myself (Department of Physics and Astronomy). We meet on a biweekly basis to plan and evaluate the course, designed as an educational experiment.

At the insistence of Candace and Nancy, we identified and developed course assessments as part of the planning process, not after the fact as is commonly done. These instruments included a concept map test, attitudinal survey, relatedness rating task, and a misconceptions measure--now called the Astronomy Diagnostic Test (ADT) versions 1.X (Zeilik et al. 1997, Paper I; Zeilik et al. 1999, Paper II). Additional expansion and development of the test resulted in the ADT version 2, which targets the typical "Astro

101" taught in the United States and was released in June 1999 (Hufnagel 2002). This version was used in the ADT National Project (Deming 2002), the first national survey of students in "Astro 101." (ADT version 2 can be found at www.flaguide.org as a Word or PDF file.) The goal of the ADT is to check the *prior knowledge* of selected astronomical concepts--those that are usually taught in K-12 (Adams & Slater 1999).

Our most serious constraint was the large size of the class, from 150 to 280 students. We realized that the form of the survey had to be multiple (forced) choice, given large enrollments at the University of New Mexico (UNM) and elsewhere. We coped with this problem by piloting the materials in a special section restricted to 40 students in fall 1992 though spring 1994. Each student kept a journal about the class. I used their comments as feedback to drop materials, add others, and revise those that "worked" from the students' viewpoint. In addition, ongoing assessments such as minute papers (see the Minute Paper Classroom Assessment Technique at www.flaguide.org for details) were included in the large classes from fall 1994 to spring 1996. Our strategy aimed at a modest type of "scale up," although for the first large class, it felt more like a "blast off"!

Figure 1 shows a concept map of the process of course reform, which took place over four semesters through the spring 1996 semester. Note the analogous structure of "Instructional Tools" and "Assessment Tools," as was intended with the parallel and integrated development of both. For details of the instructional strategies, see Papers I and II. We focus here on the students' prior knowledge, their misconceptions, and an assessment of those misconceptions, with a goal of assessing the impact of instruction on changing misconceptions to those scientifically accepted today.

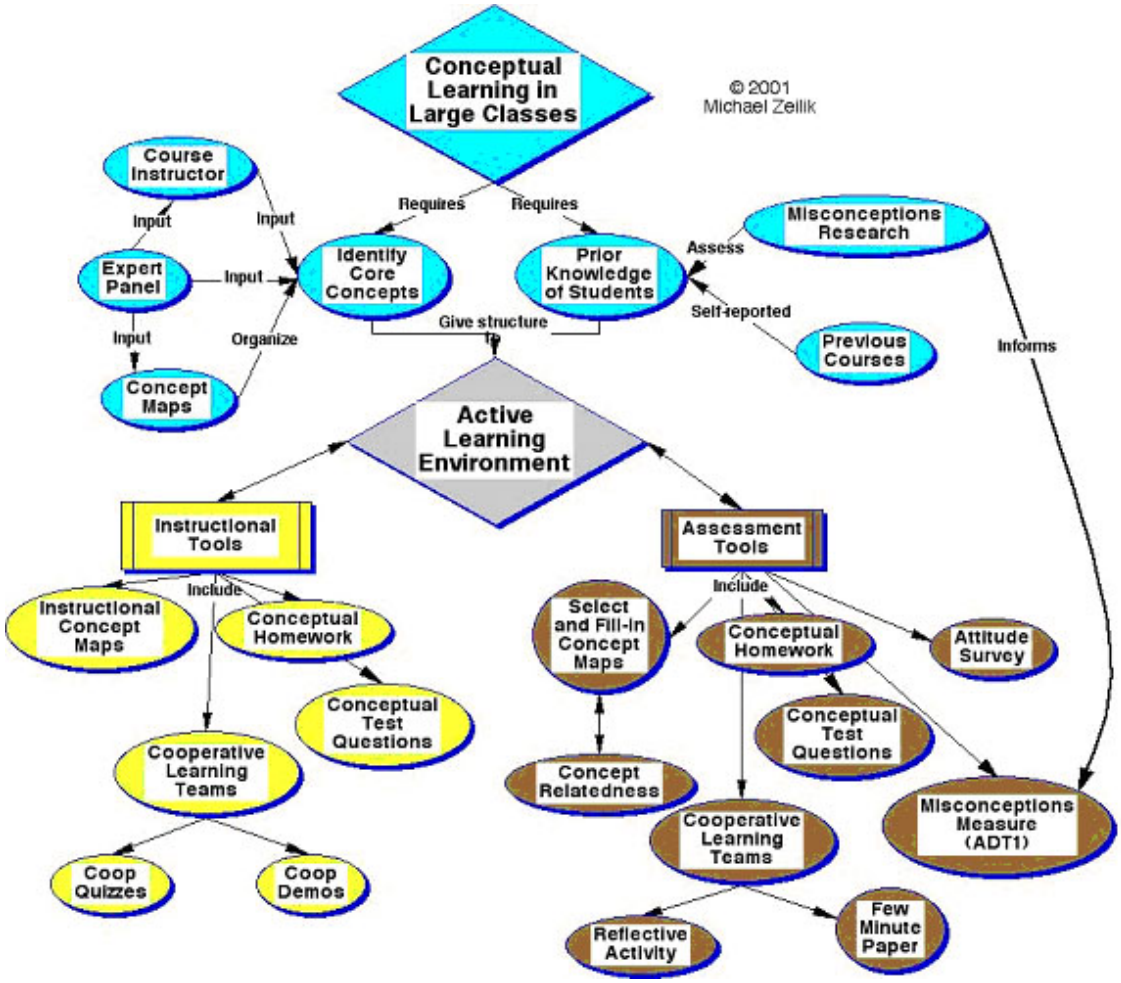

Figure 1. Concept map of experimental design

2. METHODOLOGY

We received assistance from an external expert panel of 19 experienced teachers of "Astro 101," many of whom had served on education committees of the American Astronomical Society and the American Association of Physics Teachers. A number had won teaching awards locally or nationally. The panel had two tasks: (1) to rank, out of 200 concepts, the most essential for "Astro 101"; and (2) to rate how closely these concepts were related. The goal was a community validation to ensure that any bias of mine did not determine the key concepts of the project. Once these were identified, we structured the course and the assessments around these key concepts/connections for the "big picture" of astronomy. Students would display their understanding of the material by demonstrating their connected understanding of concepts

and their application to situations similar to ones seen before (near transfer) and to novel ones (far transfer).

The cognitive sciences have identified prior knowledge as one of the robust notions that must be applied to course development (Redish 1994). This means that we must assess the students' incoming state of knowledge. To do so required a conceptual diagnostic test. We first called it a "misconceptions measure"; it is now known as the Astronomy Diagnostic Test (ADT). I examined two key sources for ideas and findings: the Project STAR assessments (long and short versions, Sadler 1992; Lightman and Sadler 1993) and the annotated bibliography of Pfundt and Duit (1994). I then reviewed my test bank to ascertain if I already had items that elucidated the misconceptions in these prior works. These items were presented to a focus group of 10-14 students who volunteered for the task for extra credit. This group met just prior to the regular class and also assisted in the development of the cooperative learning team activities. We employed focus groups for two semesters, fall 1994 and spring 1995.

For each semester at UNM, from spring 1995 to spring 1996, we used different versions of the ADT. We performed a standard item analysis for each question and calculated overall reliability of each version. Our criteria were to find items whose pre-test scores were at best 40% correct, and whose discrimination indexes were in the range of 0.2 to 0.6. (See the Multiple-Choice Test Classroom Assessment Technique at www.flaguide.org for details of item analysis.) We calculated internal reliability (technically called Cronbach's alpha); on pre-tests, it averaged 0.4--this is low, but expected on a pre-test where many students are guessing. Our post-test scores averaged 0.7, which is acceptable by item test analysis standards. An assessment must be reliable to have any use at all. In this sense, reliability is analogous to an astronomical photometer that, after repeated observations of the same star on the same telescope, gives consistent results within statistical errors.

We retained or discarded items on this basis. The result was a bank of 29 questions; 15 were used in any one class, and 14 became the core questions of the ADT versions 1.0 to 1.4 (Zeilik, Schau, & Mattern 1998). This core was the "final" version of ADT 1.X. See Table 1 for the evolutionary stages of ADT versions 1.X. Note that some subjectivity entered into the decision to keep or throw away a specific item, based on the instructor's (MZ) reaction to the pre/post results. This bias usually reflected the extent to which the concept probed by the item was considered important or was subject to specific intense instruction during the semester. (By "intense instruction," I mean that the concept had at least a cooperative team activity, a demonstration, or both.) A key goal of a reformed course is to improve instruction on key concepts and misconceptions.

Table 1. Versions of the ADT 1.X

Version	First use	Reference
1.1	Fall 1994, Astronomy 101, UNM	Paper 1 (1997)
1.2	Spring 1995, Astronomy 101, UNM	Paper 2 (1999)
1.3	Fall 1995, Astronomy 101, UNM	Paper 2 (1999)
1.4	Spring 1996, Astronomy 101, UNM	Zeilik & Bisard (2000)
1.5	Fall 1996, Astronomy 270, UNM	Zeilik & Bisard (2000)
1.6	Spring 1997, Astronomy 271, UNM	Zeilik & Morris (2002)

3. RESULTS

We were able to match pre- and post-test scores for 586 UNM participants in the project from fall 1994 to fall 1995. The mean pre-test score was 38% ± 8.2% (SD); for the post-test, 69% ± 11% (SD). These result in a normalized gain index, $\langle g \rangle$, of

$\langle g \rangle$ = (post% - pre%)/(100% - pre%) = 0.48

A $\langle g \rangle$; of 0 means no gain, while a $\langle g \rangle$ of 1 indicates that all possible gain occurred; so 0.48 means that classes gain about half of the possible gain over one semester in our reformed astronomy course. See Hake (1998) for the usefulness of the normalized gain index based on a large sample surveyed by the Force Concept Inventory (FCI).

Another useful metric used widely in the education research literature is that of effect size, which is defined as the difference between the means of the post- and pre-scores, divided by the mean of their standard deviations:

Effect size = ES = (post-test -- pre-test)/mean SD of the distributions

In other words, the effect size is the difference between the two distributions normalized by the spreads in each distribution. We found an ES = 1.90. This means that 97% of the post-scores are above the mean of the pre-scores. In educational research, an effect size above 0.8 is considered exceptional (Cohen 1998). The fact that our results involve a large sample makes the outcome robust and of great practical import: a conceptual approach significantly improves student learning.

These results cannot be compared directly to the ADT 2, which only has two items in common with ADT 1.X. We can acquire an idea of the baseline differences by comparing $\langle g \rangle$ for the earlier ADTs and ADT 2. The ADT national sample (Hufnagel et al. 2000; Deming 2002) yielded an average value of 32.4% (standard error of 0.21%) for the pre-course test, and 47.3% (standard error of 0.32%) for the post-course test. The ADT 2 is a much more difficult assessment than ADT 1.X by design; it underwent many semi-structured clinical interviews and validation of the test questions (Hufnagel 2002). Perhaps most

important, the items are written in the "natural language" of the students far more so than for ADT 1.X. This lack of astronomical jargon makes the pre-test far more accessible by novice students.

Reflecting upon some 10 years of development of the ADT, I have been impressed by the difficultly of developing a reliable and valid conceptual diagnostic test. It is not a task you can do alone in your office. You must have a team, and that team includes students. You must listen to them very carefully, both in informal and formal assessments. Because you can really attend to a small percentage of the class (say 10%), you must temper these opinions with an assessment of the class as a whole. The integration of qualitative with quantitative probes orthogonal dimensions that give a deep view of students' conceptual change with respect to misconceptions. Finally, I urge you not to change any of the ADT items. I have noted a tendency for instructors to start rewriting the questions when they first see the ADT and add astronomical terms. Resist this natural temptation!

Why go through this effort? If you utilize a validated conceptual diagnostic test, the "effort" really is not that great--about 30 minutes of class time at the start and end of a semester, and then about 30 more minutes to analyze the results. The results will likely surprise you and spur you to innovative instruction. I urge you to base your transformations on your learning outcomes united with the evidence from physics and astronomy education research.

This work was supported in part by National Science Foundation grants DUE-9253983 and 9981155.

References

Adams, J., & Slater, T. 1999, *Journal of Geophysics Education*, 48, 39.

Bisard, W., & Zeilik, M. 1998, "Restructuring a Class, Transforming the Professor: Conceptually Centered Astronomy with Actively Engaged Students," *Mercury*, 27:4, 16.

Cohen, J. 1998, *Statistical Power Analysis for the Behavioral Sciences* , Mahwah, N.J.: Lawrence Erlbaum Associates.

Deming, G. 2002, "Results from the Astronomy Diagnostic Test National Project," *Astronomy Education Review* , 1(1):52.

Hake, R. R. 1998, "Interactive-Engagement vs. Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses," *American Journal of Physics*, 66, 64.

Hufnagel, B. 2002, "Development of the Astronomy Diagnostic Test," *Astronomy Education Review*, 1(1):47.

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, http://www.atnf.csiro.au/pasa/17_2/.

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

Pfundt, H., & Duit, R. 1994, *Bibliography: Students' Alternative Frameworks and Science Education*, 4th edition, Kiel, Germany: Institute for Science Education of the University of Kiel.

Redish, E. F. 1994, "Implications of Cognitive Studies for Teaching Physics," *American Journal of Physics*, 62, 796.

Sadler, P. M. 1992, "The Initial Knowledge State of High School Astronomy Students," Dissertation, Graduate School of Education, Harvard University.

Zeilik, M., & Morris, V. 2002, "A Research-based Astronomy Course for Science, Mathematics, and Engineering Majors," submitted to the *Journal of College Science Teaching*.

Zeilik, M., Bisard, W., & Lee, C. 2002, "Research-based Reformed Astronomy: Will it Travel?," *Astronomy Education Review*, 1(1):33.

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

Zeilik, M., Schau, C., & Mattern, N. 1998, "Misconceptions and Their Change in University-level Astronomy Courses," *The Physics Teacher*, 36:12. Core version of the ADT 1.X.

Zeilik, M., Schau, C., & Mattern, N. 1999, "Conceptual Astronomy. II. Replicating Conceptual Gains, Probing Attitude Changes Across Three Semesters," *American Journal of Physics*, 67:923.

ÆR
Research and Applications