Conceptual Diagnostic Tests

Michael Zeilik
Department of Physics and Astronomy
University of New Mexico

WHY USE CONCEPTUAL DIAGNOSTIC TESTS?
To assess how well students understand key concepts in a SMET field prior to, during, and after instruction.

WHAT IS A CONCEPTUAL DIAGNOSTIC TEST?
A test with items in a multiple-choice or short-answer format that has been designed with common misconceptions in mind.

WHAT IS INVOLVED?

Instructor Preparation Time: Minimal for using available tests; moderate for designing your own questions.
Preparing Your Students: Nothing special.
Class Time: At least 30 minutes for a complete test.
Disciplines: Appropriate for all.
Class Size: Small and large.
Special Classroom/Technical Requirements: Machine scoring of scannable forms.
Individual or Group Involvement: Either.
Analyzing Results: Can be machine scored for large classes; diagnostic tests are generally not graded.
Other Things to Consider: Need to match tests to course goals.

Contents

Description

A conceptual diagnostic test aims to assess students' conceptual understanding of key ideas in a discipline, especially those that are prone to misconceptions. Hence, they are discipline-specific, rather than generic. The format typically is multiple-choice, so that a conceptual diagnostic test can be given efficiently to large numbers of students and machine scored. Unlike traditional multiple-choice items--and this is crucial!--the distractors are designed to elicit misconceptions known from the research base. (See "Theory and Research.") A student must have a clear understanding of a concept in order to select the correct response. Because conceptual diagnostic tests can be scored quickly, they can be used as formative as well as summative assessments (see the Primer).

Assessment Purposes

- To reveal the misconceptions students bring as prior knowledge to a class.
- To measure the conceptual gains of a class as a whole.
- To identify concepts that are weak areas of understanding.

Limitations

To develop reliable and valid conceptual diagnostic tests is a major, long-term undertaking. Only a limited number of such tests are currently available, and those may not match your course goals. Your field may be one in which no such tests have been developed.

Teaching Goals

- Learn concepts and terms of a subject.
- Develop higher-level thinking skills, strategies, and habits.
- Recognize common misconceptions in order to avoid or change them.

Suggestions for Use

Adopt already-developed, field-tested instruments

Well-established conceptual diagnostic tests (such as the Force Concept Inventory in physics) are: research-grounded, normed with thousands of students at diverse institutions, the product of many hours of interviews to validate distractors, and subjected to intense peer review. Individual faculty are unlikely to match such efforts. You can adopt a test, but you must follow the guidelines for its use for the results to be valid and reliable. Generally that means that you give the assessment as a pre- and post-test, secure the tests, give enough time so that all students can complete all questions, state that it is a diagnostic test and has no effect on grades, and give all items in the order presented on the instrument.

Adopt already-developed test items

You may not wish to give a complete instrument for your classroom assessment. Instead, you can give selected items from a well-developed instrument (Figure 1). While you cannot compare your results to those normed from the complete instrument, this limited use may better match your course goals.

As seen from your location, when is the Sun directly overhead at NOON (so that no shadows are cast)?

A. Every day.
B. On the day of the summer solstice.
C. On the day of the winter solstice.
D. At both of the equinoxes (spring and fall).
E. Never from the latitude of your location.

Figure 1. Sample item from the Astronomy Diagnostic Test (ADT) version 1 (Zeilik et al., 1998). The correct response is "E".

Develop your own conceptual diagnostic questions
The main advantage with this process is that you can match questions closely to your course goals. You can try out one or two questions at a time; this method will take very little class time and gives you the chance for immediate revision based on feedback from the class. Over a few semesters you can build up a bank of well-constructed items. However, you really need to investigate the research literature before you take this path.

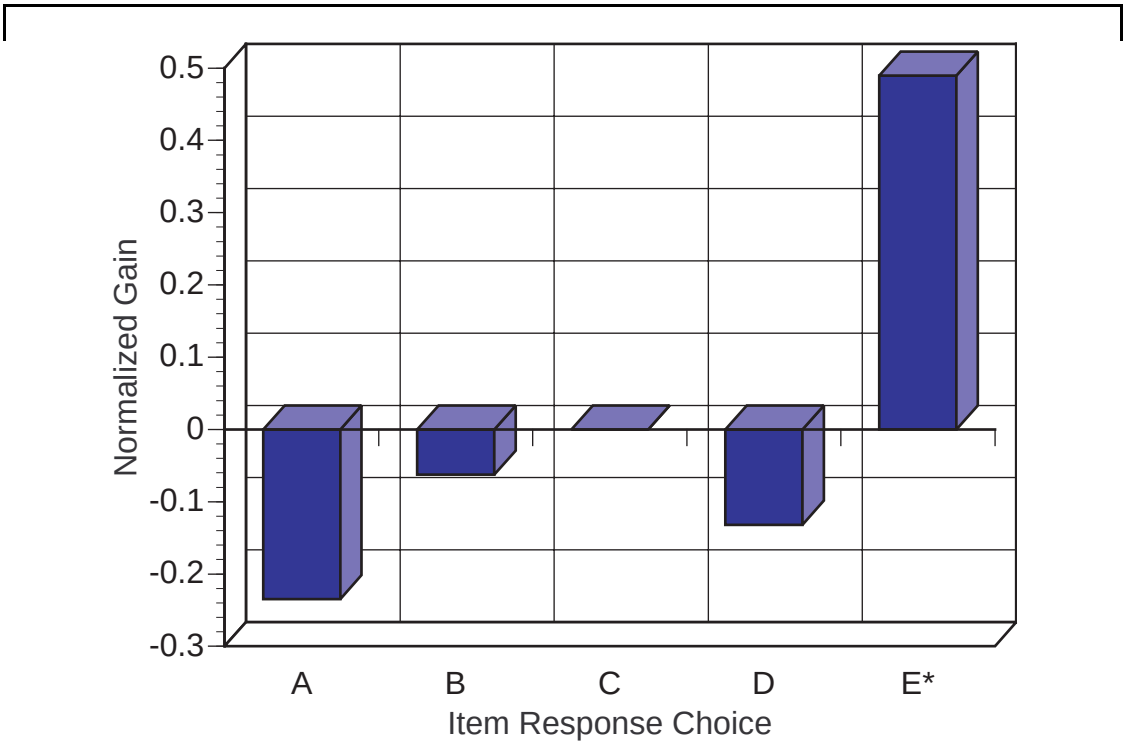
Step-by-Step

- Based on your experience or course goals, and perhaps a consensus of your colleagues, make a list of the most important concepts in your course.
- Check the misconceptions literature in your discipline to see if the research has revealed any misconceptions related to your key concepts. (See "Theory and Research.")
- If you don't find any explicit research materials, reflect on your own experience as a student and instructor. I have found, for instance, that most of the concepts I've identified as "key", my students identify as "difficult"; focus then on these.
- If you find a diagnostic test already available in your discipline, request a copy. Compare the items to your course goals and key concepts. If the test as a whole aligns with these, use it! If not, examine specific items for applicability to your course.
- Follow the developers' protocol for giving the test exactly. Contact them if you have any questions!
- Write brief, multiple-choice questions, using standard guidelines for developing good items. Avoid technical jargon; use plain English. (If you do a good job, students may perceive these questions as "hard" or "tricky" because rote memorization will not ordinarily give the correct answer.)
- Interview a few students to debug your questions. You want students to choose the "wrong" responses for the "right" reasons--that is, a certain misconception or a poor line of reasoning that leads them astray. Alternatively, debug the questions with the whole class, as described in the next section.
- The best use of a diagnostic test is as a pre/post assessment. You do not have to wait until the end of the semester to give a post test; you can give it right after instruction on a coherent instructional unit. If possible, you should obtain a standard item analysis, so you can check for problems with the test as a whole or with individual items. (On most campuses, this analysis is provided by the computer center.)

- One way to quantify the pre/post gains is to calculate a **gain index** (Hake, 1998). This is the actual gain (in percentage) divided by the total possible gain (also in percentage). Hence, the gain index can range from zero (no gain) to 1 (greatest possible gain). This method of calculating the gains normalizes the index, so that you can compare gains of different groups and classes even if their pretest scores differ widely. (Note that it is also possible to get negative results!) The formula is

  gain index = (%post - %pre)/(100 - %pre)

- You can do this gain calculation in two ways: (1) find the gain for the average pretest and average posttest score of the class as a whole (gain of the averages); or (2) average each student's gain (average of the gains). If your class size is greater than about 20 to 30, these two techniques will give essentially the same result. For the item in Figure 1, the pretest score (spring 1995) was 23%, the posttest score was 64%, so the gain of the averages was 0.53. You can also calculate the gain index for each response (Figure 2) and that way see how students changed their responses from pre to post. Why do this calculation? It gives you a "one-number" value so that you can compare classes over time (summative) or on-line (formative).



Variations

Give individual items to pairs and/or cooperative learning teams for interactive discussion
Ask students to form pairs with their neighbors or form up into cooperative learning teams. Present a conceptual question (can be done by handouts or with an overhead projector). Request them to think about it for a minute. Then poll the class for their chosen answers (this is the advantage to a multiple-choice format). In a small class, people can just raise their hands. In a large one, however, students will look around waiting to see if others are raising their hands so they can "vote' with the majority. To avoid this problem, and to help you tally, give each students large flash cards with numbers/letters on one side. Students hold these up all at the same time. You can get a quick sense of answer choices by scanning the responses. Then ask the students to discuss their choices with their partner or learning team for a minute or two. Poll the class again; the responses should move toward the correct one. If not, do a review on the spot by asking a few students to explain their responses.

If you want to "capture the data", request that each student take a piece of paper and draw a horizontal line across the middle. Above the line they write their individual choice of answer and a brief (one sentence!) explanation. Below it, they write their choice and explanation *after* the discussion with their partner or group.
This procedure is a great way to "debug" your own conceptual questions!
Note: This variation has close similarities with ConcepTests; see that CAT.

Individual or Group Interviews
To use interviews for a deeper diagnostic probe, you need to establish a protocol to follow for a consistent structure. Individual interviews are very time-consuming to conduct and analyze. It is important that you choose your sample of students well to be sure that it is representative of the class. Group interviews probe a wider range of thinking more efficiently. In both cases, you may need a second person to record the interview while you conduct it. Taping and transcribing are really research requirements, not an assessment ones. You may just want qualitative insights here, not quantitative ones. See the CAT on Interviews.

Two-Tier Questions
As a deeper probe in a less time-consuming format than interviews, you can ask the students to respond to test questions at two levels (Treagust, 1988). First they select and answer (tier one), and then they write a justification (tier two). The second step may be impractical for a large class; you can use cooperative learning teams to promote the process. A third step that works well during interviews is to ask the student "how confident" she is about her response and explanation.

Analysis

Your first objective, especially for formative assessment, is to get a "feel" for your students' misconceptions, conceptual change, and the impact of your instruction (positive or negative!). A few individual interviews or a well-chosen group interview will provide insights, but beware of limited reliability. In-class pair or group interactions are more reliable, and you can explore a good number of concepts quickly (each item will take only about 5 minutes). Your second objective, after you've given written tests, can be more quantitative, such as calculation of a gain index. If you follow a protocol, you can compare your results to those from other classes/institutions and perhaps even provide data for research.

Do *not* use the results of diagnostic tests to assign grades! That is a misuse of the technique and the tools. For individuals, you can use the results as a guide to tutor students in weak areas. Or you can use the pretest results to help you assign people to cooperative learning teams (I mix my teams with people who have high, low, and middle scores on the diagnostic test).

Pros and Cons

- Field-tested diagnostic instruments are research-based and have undergone extensive scrutiny for their validity and reliability; they have been used in a variety of settings.
- You can use scores to make up heterogeneous cooperative learning teams.
- Such tests are short, involve limited class time, and are easy to score.
- These tests are extremely useful for formative and summative assessments over semesters.
- If you follow the protocol for a given instrument, you can tap into a large comparative database.
- Short, in-class applications are quite revealing in terms of the impact of various facets of instruction.
- You can develop your own tests, specifically tailored to your course goals.

However:

- Very few field-tested instruments are currently available; their items may not match your course goals.
- To use these diagnostic tests properly often requires that you follow a formal protocol and keep the tests secure.
- Limited use of selected items may not be possible.
- You must be careful not to misinterpret the results.
- Designing your own tests is extremely demanding and will take at least a few semesters for each course; you must do interviews; reliability and validity may be hard to establish.
- Your students will declare such questions to be "hard" or "tricky" until they realize that you *really* mean they are "diagnostic"!

Theory and Research

Effective conceptual diagnostic tests are grounded in the research on alternative conceptions in science, commonly perceived as misconceptions by scientists (see the extensive references in Wandersee, Mintzes, and Novak, 1994). The latest bibliography of research papers (Pfundt and Duit, 1994) contains some 3600 entries, of which 66% are related to physics, 20% to biology, and 14% to chemistry. The bulk of the physics work has been in classical mechanics and electricity.

Many of the classic papers in misconceptions research deal with students younger than those in universities. But they are not irrelevant to a higher education context! You will discover that a large percentage of your students hold these alternative frameworks about the workings of the natural world. I have presented a separate list of sources to a selected few disciplinary-specific papers.

Force Concept Inventory

The *Force Concept Inventory* is the best developed and most widely used diagnostic test in physics (Hestenes, Wells, and Swackhamer, 1992; Hake 1998). Ibraham Halloun, Richard Hake, Eugene Mosca, and David Hestenes revised this test in 1995; this revision is the current version. The FCI has 30 qualitative items, with subscales, dealing only with the Newtonian concept of force. It is extremely effective in eliciting the "commonsense" notions of students about motion. The questions were designed to be meaningful to students without formal training in mechanics.

Mechanics Baseline Test
The *Mechanics Baseline Test* is more difficult than the FCI (Hestenes and Malcolm, 1992). It focuses on concepts that can be understood only *after* formal training in Newtonian mechanics. It contains 26 items, some of which involve simple calculations (but none that require a calculator). I have found the MBI challenging even for new graduate students in physics, who would be expected to do well on such a test.

Conceptual Surveys in Physics
The Two-Year College Physics Workshop is developing a set of conceptual surveys: *Conceptual Survey in Electricity* (CSE), *Conceptual Survey in Magnetism* (CSM), and *Conceptual Survey in Electricity and Magnetism* (CSEM). The CSEM is a shorter, combined subset of the CSE and CSM. The goal is to provide conceptual tests for common physics topics other than mechanics. The latest versions are Form G (7/98). The CSE has 32 items, CSM 21, and CSEM 32.

Conceptual Learning Assessments for Workshop Physics
Three assessment tests have been used to assess conceptual learning gains in Workshop Physics courses. They are: *Force and Motion*, *Heat and Temperature*, and *Electric Circuits*. They are available with a password to educators as Microsoft Word files.

Astronomy Diagnostic Test
The *Astronomy Diagnostic Test* (version 2) originated as an assessment for the conceptual astronomy research project (Zeilik et al., 1997). Early versions relied heavily on the results of an assessment from Project STAR; this test contained 60 physics and astronomy items. Lightman and Sadler (1993) give a 16-item version of this STAR assessment. Zeilik, Schau, and Mattern (1998) presented 15 central items from ADT (version 1). Although this ADT version was subjected to small-group interviews, it was never validated by extensive individual interviews. These interviews have been carried out at Montana State University (Adams and Slater) and at the University of Maryland (Hufnagel and Deming) and formed the basis for ADT (version 2).

California Chemistry Diagnostic Test (CCDT)
This 44-item test was developed by a 16-member team in response to a need in California for a "Freshman Chemistry Placement Test"to be used at all levels of institutions (Russell, 1994). The CCDT was given a trial run in 1987 in 53 schools with over six thousand students; a revised version was tested in 1988. The correlation coefficient with final grades in first term general chemistry was 0.42. The CCDT is availble through the American Chemical Society's Division of Chemical Education: ACS DivCHED Examinations Institute, 223 Brackett Hall, Clemson University, Clemson SC 29634-1913.

Links

The American Association of Physics Teachers Web site <http://www.aapt.org> provides access to some of the conceptual learning assessments. Click on PSRC to link to the "Physical Science Resource Center" at http://www.prsc-online.org. Then click on "Evaluation Instruments" to find "Conceptual Learning Assessments."

Copies of the Conceptual Surveys are available by contacting Curtis Hieggelke at curth@jjc.cc.il.us. Please give your complete mailing address.

The revised **Force Concept Inventory (FCI)** (I. Halloun, R.R. Hake, E.P. Mosca, and D. Hestenes), the B (Hestenes & Wells) Survey are available to authorized educators (as .PDF files) from the Workshop Modeling Project (http://modeling.la.asu.edu/modeling/R&E/Research.html) at Arizona State University.

Three assessments in the Workshop Physics Courses. Each of these assessments with answer sheets and answer keys are available to authorized educators (as Microsoft Word files) from the Workshop Physics project (http://physics.dickinson.edu/PhysicsPages/Workshop_Physics/Instructor_Resources/Curricular_Materials/Conceptual_Assessments) at Dickinson College.

Information on the American Chemical Society's exam materials (CCDT and many others) can be found at HREF= "http://tigerched.clemson.edu/exams".

Sources
Misconceptions research

General

- Driver, R. (1993). The pupil as scientist? London: Milton Keynes.
- Pfundt, H. and Duit, R. (1994). Bibliography: Students' Alternative Frameworks and Science Education, 4th edition, Kiel: Germany.

Astronomy

- Nussbaum, J. (1979). Children's conception of the earth as a cosmic body: A cross-age study. Science Education, 63, 83-93.
- Sneider, C. and Pulos S. (1983). Children's cosmographies: Understanding the earth's shape and gravity. Science Education, 67, 205-221.
- Vosniadou, S. (1990). Conceptual development in astronomy. In S. Glynn, R. Yeany, and B. Britton (eds.), The psychology of learning science (pp. 149-177). Hillsdale, NJ: Lawrence Erlbaum.

Biology

- Arnaudin, M. W. and Mintzes, J. J. (1985). Students' alternative conceptions of the circulatory system: A cross-age study. Science Education, 69, 721-733.
- Bell, B. (1981). When is an animal not an animal? Journal of Biological Education, 15, 213-218.
- Wandersee, J. H. (1986). Can the history of science help science educators anticipate students' misconceptions? Journal of Research in Science Teaching, 23, 581-597.

Chemistry

- Ben-Zvi, N. and Gai, R. (1994). Macro- and micro-chemical comprehension of real work phenomena. Journal of Chemical Education, 71, 730-732.
- Hackling, M. and Garnett, D. (1985). Misconceptions of chemical equilibia. European Journal of Science Education, 7, 205-214.
- Nakhleh, M. B. (1992). Why some students don't learn chemistry: Chemical misconceptions. Journal of Chemical Education, 69, 191-196.
- Novik, S. and Menis, J. (1976). A study of student perceptions of the mole concept. Journal of Chemical Education, 53, 720-722.
- Stavy, R. (1988). Children's conception of gas. International Journal of Science Education, 10, 553-560.

Physics

- Champagne, A., Klopfer, L. and Anderson, J. (1980). Factors influencing the learning of classical mechanics. American Journal of Physics, 48, 1074-1079.
- Clement, J. (1982). Studies of preconceptions in introductory mechanics. American Journal of Physics, 50, 66-71.
- Fredette, N. and Clement, J. (1981). Student misconceptions of an electric current: What do they mean? Journal of College Science Teaching, 10, 280-285.
- Watts, D. M. (1985). Students' conceptions of light—A case study. Physics Education, 20, 183-187.

Diagnostic tests

- Bisard, Walter and Zeilik, Michael (1998). Conceptually centered astronomy with actively engaged students. Mercury, 27 (4), 16-19.
- Hake, Richard R. (1998). Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics, 66 (1), 64-74.
- Hestenes, David and Wells, Malcolm (1992). A mechanics baseline test. The Physics Teacher, 30, 159-166.
- Hestenes, David, Wells, Malcolm, and Swackhamer, Gregg (1992). Force concept inventory. The Physics Teacher, 30 (3): 141-151.
- Lightman, Alan and Sadler, Philip (1993). Teacher predictions versus actual student gains. The Physics Teacher, 31 (3): 162-167.
- Odom, A. L. and Barrow, L. H. (1995). Development and application of a two-tier diagnostic test measuring college biology students' understanding of diffusion and osmosis after a course of instruction. Journal of Research in Science Teaching, 32 (1): 45-61.
- Russell, A. A. (1994). A rationally designed general chemistry diagnostic test. Journal of Chemical Education, 71 (4): 314-317.
- Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. International Journal of Science Education, 10 (2), 159-169.
- Wandersee, James H., Mintzes, Joel J., and Novak, Joseph D. (1994). Research on alternative conceptions in science. Handbook of Research of Science Teaching and Learning, edited by Dorothy L. Gabel. New York: Macmillan Publishing, 177-210.
- Zeilik, Michael, Schau, Candace, and Mattern, Nancy (1998). Misconceptions and their change in university-level astronomy courses. The Physics Teacher, 36: 104-107.
- Zeilik, M., Schau, C., Mattern, N., Hall, S., Teague, K., and Bisard, W. (1997). Conceptual astronomy: A novel model for teaching postsecondary science courses. American Journal of Physics, 65 (10): 987-996.

Mike's Story

In 1992, I received a grant from the National Science Foundation to transform the introductory astronomy course for non-science majors. Little did I know then that the effort would transform my professional life! My proposal imagined an "electronic collaboratory" for the class. In the pilot tests, it was a total failure.

Meanwhile, the grant forced me to review the literature on effective teaching strategies. I came across the Force Concept Inventory in physics. The idea clicked: For years I had thought that the real measure of the effectiveness of a course was the gains of the students, from their "initial condition" at entry to their "final state" at the end. But I never could work out an implementation of the idea. Luckily, Candace Schau was a member of my interdisciplinary research team and brought her expertise in statistics and assessment. She convinced me that we needed a "misconceptions measure" for a pre- and post-test. We developed this assessment based on the misconceptions research in astronomy (especially that of Project STAR). I found it a difficult task, but it did illuminate patterns I had perceived in my 20+ years of teaching the introductory astronomy course. That clinched it for me.

We labored on the Astronomy Diagnostic Test (ADT) for four semesters. I did not realize until a year after the final effort that we were developing the astronomical version of something like the Force Concept Inventory, though not as highly focused as the FCI. As I gave talks about the conceptual astronomy project, people asked me for copies of the test. In summer 1998, the Astronomical Society of the Pacific (ASP) met in Albuquerque; the meeting included a symposium on the introductory astronomy course. I announced the availability of ADT 1.0 and handed out about 100 copies. Other requests came in by email. During the ASP meeting, Jeff Adams, Tim Slater, Beth Hufnagel and others agreed that we needed an ADT version 2, based on version 1, but revised in the light of student interviews. These were conducted in fall 1998 at Montana State University and the University of Maryland. These collaborative efforts resulted in ADT version 2.