Using diagnostic test items to assess conceptual understanding of basic biology ideas:

A plan for programmatic assessment

Kathy S. Williams, Kathleen Fisher, Dianne Anderson, Mike Smith

Introduction

University departments and programs are being encouraged to engage in systematic programmatic assessment, but little guidance is provided as to how to best achieve this goal. Assessing the effectiveness of an undergraduate program of study is different from assessing performance of individual students or faculty. What are the best indicators of success? What can provide the best insights into strategies for improving the program?

It is widely known that students come to the university with many naive conceptions about scientific topics, conceptions that are at odds with established scientific conceptions. Thus, one approach for evaluating science programs is to ask, “To what extent have the students, enrolled in a particular course of study, given up their naive preconceptions and committed to accepted scientific conceptions?” Science is often counter-intuitive and it can be difficult for an individual to let go of a strongly-held preconception and acquire a corresponding idea that is scientifically sound. Further, there is evidence that one idea is not readily replaced by the other, but rather that the two ideas (‘naïve’ and ‘sophisticated’) often exist side by side in an individual’s mental structure, each competing with the other (citation). One version may be recalled in the classroom, while the other is employed in other settings. Eventually, one of those competing ideas gains dominance and the other fades into disuse.

When the new idea gradually gains dominance, this extended process is known as conceptual change. On the other hand, sometimes a new idea is retained just as long as the class in which it was introduced, and then it fades away again. This is known as short-term or temporary learning. Research indicates that lecture teaching is a relatively ineffective means for producing conceptual change, whereas engaging students in hands-on problem-solving, especially with appropriate hands-on, minds-on experiences, can be much more effective. This is especially true when the activities are combined with peer collaboration and discussion.

Here we define “diagnostic tests” as multiple choice tests that have been developed to assess the extent to which students choose scientifically sound ideas over commonly held preconceptions. Diagnostic test items are often two-tiered, wherein the first stem or item asks about a scientific ‘fact’ and the second stem asks about the respondent’s reason for choosing a particular response option in the first item. The second item typically offers four response options, one of which is scientifically correct and three of which represent common alternative conceptions. Other diagnostic tests present an authentic scenario in a short paragraph and pose questions about it, also with each response option aligned with either the scientifically correct or alternative conceptions. Since the alternative conceptions are attractive to many test-takers, scores on diagnostic tests are considerably lower than scores on traditional, content-based tests. Student performance on diagnostic tests also changes slowly as a function of time in an academic program, as you will see below. The stability of test scores over time is one of the features that makes these tests attractive for programmatic assessment.

Since the instructional program in biology at our institution relies heavily on traditional lecture and lab formats, and since diagnostic tests provide a relatively stable measure of gains, and since diagnostic tests aim to assess deep understanding of important ideas (as opposed to more superficial knowledge of vocabulary), we feel that diagnostic tests can provide an effective measure of learning gains among our biology majors. Thus, we chose diagnostic testing as one indicator of programmatic assessment. Other indicators may be added subsequently. This paper provides an overview of our progress to date.

Description of Diagnostic Tests

If we wanted to study the process of conceptual change, we would need to examine that process as it occurs in individual students, observe the timing and nuances of the changes, and explore what specifically triggered the changes. But that is not our purpose. Instead, we are using measures of conceptual change to assess the effectiveness of our undergraduate biology program.

We are developing five diagnostic tests to assess understanding of: Natural Selection, Osmosis & Diffusion, Cell Division (processes of mitosis & meiosis), Energy & Matter Transformations, and Nature of Science. The first two tests are in final form, but the last three are still undergoing revision and refinement.

Natural Selection Diagnostic Test. Items were selected from the Conceptual Inventory of Natural Selection ('CINS', Anderson, Fisher & Norman, 2002), with minor modifications (with permission from the authors). Included are five questions about the Galapagos finches and five about the Venezuelan guppies. This test is now stable.

By 'stability, we mean that the diagnostic test has been evaluated with several classes and meets our expectations that a) all incorrect responses are selected by some students, b) modest gains in proportion of students selecting the scientifically correct response are seen as students progress through the major, and c) the proportion of students selecting the correct response falls within the range of 30% to 70% (Kaplan & Saccuzzo, 1997). We are working on a fourth criterion at the present time, as described below, by conducting interviews to determine if students are interpreting the items and responses in the ways that are intended.

Osmosis & Diffusion Diagnostic Test. Items were selected from the Osmosis and Diffusion test (Odoms & Barrow 1995), with modifications (with permission from the authors). This diagnostic test consists of ten two-part questions, twenty questions in all. This test is now also stable.

Cell Division (mitosis & meiosis) Diagnostic Test. This test is being developed entirely by this team and is still under development, as we consider special issues related to this topic. A diagnostic test aims to identify students' understanding of ideas as opposed to their memorization of biological terms. Thus, it is desirable to avoid using biological terminology or jargon in the test items. This is quite challenging in most areas of genetics including mitosis and meiosis. Consider the following item:

1. A human egg and sperm join together to form the beginnings of a brand new baby. The fertilized cell prepares to divide. Prior to the first division of the fertilized call, each maternal chromosome
   a. is copied precisely
   b. pairs with the corresponding paternal chromosome
   c. undergoes crossing over
   d. all of the above
   e. (a) and (b)

The first two sentences, '*A human egg and sperm join together to form the beginnings of a brand new baby. The fertilized cell prepares to divide,'* is intended to convey the fact that we are looking at cell division in a diploid somatic cell, and that would involve mitosis. But mention of *sperm* and *egg* seems to trigger thoughts of meiosis, and most students answer this item incorrectly. To prompt deeper thought about the nature of the cell, we may break the item into two parts as follows, creating a 3-part question series overall

1. A human egg and sperm join together to form the beginnings of a brand new baby. The resulting fertilized egg is a ...
   a. haploid somatic cell
   b. diploid somatic cell.
   c. haploid germ cell
   d. diploid germ cell.

2. As an embryo begins to form and its cells divide, each maternal chromosome
   a. is copied precisely
   b. pairs with the corresponding paternal chromosome
   c. undergoes crossing over
   d. all of the above
   e. (a) and (b)

3. The reason for my response is that
   a. maternal chromosomes typically pair with similar paternal chromosomes before a cell divides.
   b. crossing over assures a healthy degree of genetic variability.
   c. somatic (body) cell division generally reproduces each chromosome exactly.
   d. genetic variation is essential for species survival.
   e. all of these events occur in cell division.

This is the approach we have taken so far, focusing on the process of mitosis and meiosis. Our Biology colleagues now would like us to expand the array of diagnostic tests to address a larger view of cell division (including regulation), development, and molecular biology, and we are embarking on that next.

The *Energy & Matter Transformation Diagnostic Test* has been created on the basis of student short essay responses and published research on prevailing alternative conceptions (D. Ebert-May citation). Generally this test consists of tiered items, in which one item poses a question and the next item asks for the respondent's reasoning. We are planning student interviews to help complete the conception specification table and add additional question items.

The *Nature of Science Diagnostic Test* we are using now asks for simple Agree/Disagree responses. In all cases, items are designed to help us learn how our students are thinking about specific essential concepts. We are leaning toward developing a more focused *Nature of Biology* diagnostic test with more response options that are tied to more alternative conceptions. It is likely that we will need to interview students and ask them to answer open-ended questions to help with that too.

Description of Programs Being Assessed

The Biology Department offers a Biology B.S. with emphases in a) Cellular and Molecular Biology, b) Ecology, c) Evolution and Systematics, d) Marine Biology, and e) Zoology. Students may also earn a B.A. in Biology, or a B.S. in Microbiology. All programs are exactly the same in their lower division science course requirements, and all require the same 3 “core” upper division courses in ecology, evolution & genetics, and biochemistry, cell, & molecular biology. About half of all majors also take an upper division microbiology course as their required organismal biology course. We chose to evaluate knowledge gains in those 4 upper division courses.

Methods

As part of the validation process, the diagnostic tests are administered as pre- and post-course tests in both major and non-major biology courses at our large public university and at adjacent community colleges. Since classes at all participating schools are large, the students in each class are divided into four or five groups. Each group receives a different diagnostic pre-test, and then gets a post-test on the same topic.

Description of Diagnostic Test Development

The diagnostic test items focus on biology ideas that are either *basic* (e.g., osmosis & diffusion) or *big* (e.g., natural selection), meaning they are topics that serve as either foundational knowledge or over-arching frameworks that influence t learning of many ideas in biology. These topics are rarely being taught directly in the particular course in which the student is enrolled. Each test consists of about ten 2-part items or twenty 1-part items. As noted above, students are not graded on their performance. However, many instructors give students some fixed number of points as an incentive for taking the diagnostic tests.

We began by identifying known misconceptions (including both published findings and observations collected by the authors, especially via short answer essays). Next, the particular topics to be tested were identified. In two cases we began with existing diagnostic tests, as noted above. A specification table was created summarizing the topics of interest, common alternative conceptions, and the correct or scientific ideas (Table 1). We then worked to generate two items per topic. These are have been evaluated and refined, usually over the course of 2-3 semesters. These are our guidelines for test item creation:

1) In order to distinguish between deep level understanding of an idea versus memorization of the meaning of a particular term, we systematically avoid using biology jargon. To the best of our ability, questions are phrased in simple English. This is particularly challenging where genetics issues come into play, since there really are not commonly used terms for such ideas as gene, allele, and DNA.
2) The correct response should be unambiguous.
3) Each distracter (incorrect response) should reflect a common naïve conception and should be attractive to at least some students

Once a test is created, it is given to several expert reviewers to determine if all agree that there is one correct response. It is also given to a pool of students to determine if the items are clear and easily interpreted.

Description of Interviews [to be added]

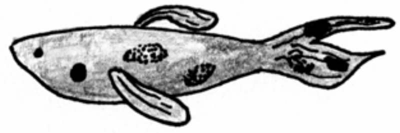
Table 1. Specification table for natural selection. Numbers in parentheses refer to corresponding item numbers.

Topic/Issue	Common Confusion or Naïve Conception (subtest item/stem #, option A-D)	Scientific Ideas (subtest item/stem #, option A-D)
A - Population stability	• All populations grow in size over time (1A, 7B) • Populations decrease (1D, 7C) • Populations always fluctuate widely/ randomly (1C, 7D)	• Most populations are normally stable in size except for seasonal fluctuations (1B, 7A)
B - Origin of Variation	• Mutations are adaptive responses to specific environmental agents (3C, 9D) • Mutations are intentional: an organism tries, needs, or wants to change genetically (3A, 3D, 9A, 9B)	• Random mutations and sexual reproduction produce variations; while many are harmful or of no consequence, a few are beneficial in some environments (3B, 9C)
C - Variation inheritable	• When a trait (organ) is no longer beneficial for survival, the offspring will not inherit the trait (4B) • Traits acquired during an organism's lifetime will be inherited by offspring (4A) • Traits that are positively influenced by the environment will be inherited by offspring (4D)	• Much variation is heritable (4C)
D - Origin of Species	• Organisms can intentionally become new species over time (an organism tries, wants, or needs to become a new species) (5C, 5D, 10A, 10C) • Speciation is a hypothetical idea (5B)	• An isolated population may change so much over time that it becomes a new species (5A, 10B)
E - Change in Population	• Changes in a population occur through a gradual change in all members of a population (2A, 8A) • Environment causes mutations to help individuals survive and reproduce (2D, 8C) • Mutations occur to meet the needs of the population (2C, 8D) • Learned behaviors are inherited (8C)	• The unequal ability of individuals to survive and reproduce will lead to gradual change in a population, with the proportion of individuals with favorable characteristics accumulating over the generations (2B, 8B)
F - Variation within a population	• All members of a population are nearly identical (6A) • Variations only affect outward appearance, don't influence survival (6B, 6C)	• Individuals of a population vary extensively in their characteristics (6D)

Any scientific data used in a test item is drawn from an actual scientific study. We feel this is important to keep our representations of scientific research 'authentic.' It is far too easy to misrepresent science when using hypothetical examples, and to thus risk introducing new misconceptions to students. This is illustrated in the following example about natural selection (Anderson et al. 2002), which briefly summarizes a study of guppies in South American streams.

Figure 1. Example of an item associated with a header describing a scientific study of guppies in South American streams. Several test items are linked to a single header.

Venezuelan Guppies



Guppies are small fish found in streams in Venezuela. Male guppies are brightly colored, with black, red, blue and iridescent (reflective) spots. Males cannot be too brightly colored or they will be seen and consumed by predators, but if they are too plain, females will choose other males. Natural selection and sexual selection push in opposite directions. When a guppy population lives in a stream in the absence of predators, the proportion of males that are bright and flashy increases in the population. If a few aggressive predators are added to the same stream, the proportion of bright-colored males decreases within about five months (3-4 generations). The effects of predators on guppy coloration have been studied in artificial ponds with mild, aggressive, and no predators, and by similar manipulations of natural stream environments (Endler, 1980).

Choose the one answer that best reflects how an evolutionary biologist would answer.

6. A typical natural population of guppies consists of hundreds of guppies. Which statement best describes the guppies of a single species in an isolated population?
   a. The guppies are identical to each other in all ways.
   b. The guppies share all of the essential characteristics of the species; the minor variations they possess don't affect survival or reproduction.
   c. The guppies are identical on the inside, but have many differences in appearance.
   d. The guppies share many essential characteristics, but also vary in many features. *

When two-tiered items are employed, the first tier (e.g., What happens when ___?) typically has two, and sometimes three, responses. In the second tier, the student is asked to give the reason for their response chosen in the first tier. Thus, the possible reason options\s given in the second tier should be equally divided between or applicable to the two initial response choices. An example from the osmosis and diffusion test is shown below.

Figure 2. Example of a two-tiered question in osmosis and diffusion.

5. If a small amount of salt (1 tsp) is added to a large container of water (1 gal or 2 liters) and allowed to set for several days without stirring, the salt molecules will
   a. be more concentrated at the bottom of the container.
   b. be evenly distributed throughout the container.
6. The reason for my answer is because
   a. there is movement of particles from a high to low concentration.
   b. the sugar is heavier than water and will sink.
   c. salt dissolves poorly or not at all in water.
   d. there will be more time for settling.

Mitosis and meiosis present special problems in trying to avoid use of jargon. Consider the example item below (Figure 3). We have avoided using the term 'somatic cell' or the more ambiguous term, 'body cell.' Instead we have talked about a cell in context. The reader has to interpret the events and figure out that this is a diploid cell now initiating the process of growth. Beginning with the terms *egg* and *sperm*, however, is likely to be misleading to the non-thoughtful reader. Given that a large proportion of upper division biology students seem to think that chromosomes pair in both mitosis and meiosis, there will be multiple reasons for incorrect responses.

Figure 3. An example item from the cell division diagnostic test.

1. A human egg and sperm join together to form the beginnings of a brand new baby. The fertilized cell prepares to divide. Prior to the first division of the fertilized call, each maternal chromosome
    - a. is copied precisely
    - b. pairs with the corresponding paternal chromosome
    - c. undergoes crossing over
    - d. all of the above
    - e. (a) and (b)
2. The reason for my response is that
    - a. maternal chromosomes typically pair with similar paternal chromosomes before a cell divides
    - b. crossing over assures a healthy degree of genetic variability
    - c. somatic (body) cell division generally reproduces each chromosome exactly
    - d. genetic variation is essential for species survival
    - e. maternal and paternal chromosomes are strongly attracted to each other

Diagnostic Test Administration

Once a test has been developed, it is administered in our “assessment” classes (lower division non-majors, and upper division majors core courses). Each item is then evaluated and refined as necessary. Any distracters that draw few or no responses are removed and replaced.

Assessments are administered as ungraded pre- and post-tests in all courses. In addition, we have some data from lower division non-major and major courses including general biology for biology majors and non-majors, collected at SDSU and local community colleges.

Since the relevant classes are large, we divide the students into subgroups, and each subgroup receives a different subtest. Thus, students are answering questions on topics not necessarily closely related to the topics covered the course. This is because the Biology Department is interested in learning about the basic capacities and knowledge of our biology students, in areas including biology, chemistry, math, and physics, to help us improve the overall biology curriculum, . Our goal is to analyze results using covariates indicating such things as which courses students previously completed, campus where lower division preparation was completed, and learning methods used in the prior courses as well as in their “current” course.

Initially, the diagnostic tests were deployed on paper and graded using ParScore, but soon after, we began using web-based tools. At first we used a university-based survey service, since our course management system (Blackboard) was not adequate at that time. Data were collected and evaluated for each test item using point bi-serial and difficulty values. The latest version of Blackboard now will produce a dataset of item responses that will let us conduct item analyses.

Evaluation and Analysis

Discriminability (point biserial values) and difficulty values are determined for all of the test items. The point biserial values indicate the ability of an individual item to discriminate between high and low performers on the entire test. The closer the point biserial value is to 1.00, the greater the discriminating power. Good test items generally result in point biserial values of between 0.30 and 0.70 (Kaplan & Saccuzzo, 1997). However, because the diagnostic tests are criterion-referenced test designed to identify concepts that students do or do not understand, rather than to discriminate among students, the point biserial values are of decreased usefulness (Gronlund, 1993).

The *difficulty* is determined by the proportion of students who respond to the item correctly.

The *reliability* of the test relates to the consistency of responses. A test must be shown to be reliable in order to be a valuable tool. As a measure of general internal consistency, we use the Kuder-Richardson 20. This method simultaneously considers all possible ways of splitting the test, so it improves on other methods of determining reliability in which the test is used only once (as opposed to test/retest methods). A good classroom test should have a reliability coefficient of 0.60 or higher (Gronlund, 1993).

Biology majors do not take their upper division courses in a specified order. For this reason, our current goal is to analyze results from each course using covariate analyses to determine the impact of such things as how many courses and which courses students previously completed, the campus where lower division preparation was completed, and the learning methods used in prior courses as well as in their "current" course.

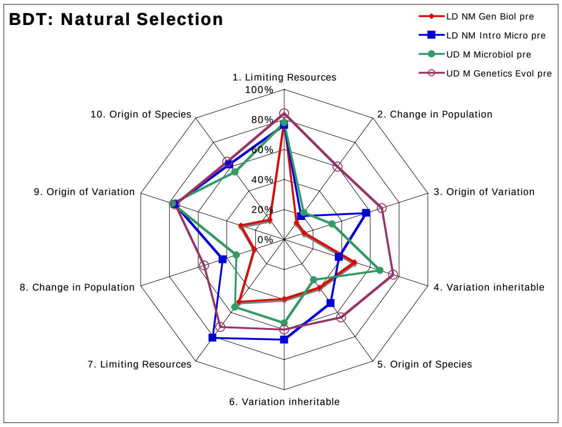
Results

Recent results demonstrate the striking consistency with which students are attracted to particular non-scientific conceptions (or "misconceptions"). Results from two of the subtests illustrate this.

Natural Selection Pretest Scores (Figure 3). The radar graph displays mean class scores on a 10-item test about natural selection, with zero in the center and 100% on the outer circle. The red (triangle) and blue (square) lines show scores for lower division biology classes, which occur in sequence. The green (closed circle), lavender (open circle) lines shows scores for upper division classes; students take these classes in any order.

Students in upper division Genetics and Evolution (Figure 3.) more accurately to questions 2, 3, 4, and 5, which ask about the origin and inheritance of traits and changes in traits over time (adaptation). Questions 2 and 8 are the most difficult for most students. They ask: what is the best way to characterize how a population changes over time, with #2 asking about finches and #8 about guppies (after Anderson et al. 2002). The questions illustrate the difficulties students have in recognizing that the *proportions* of organisms with different traits change in populations over time. Rather, they think that populations or individuals change.

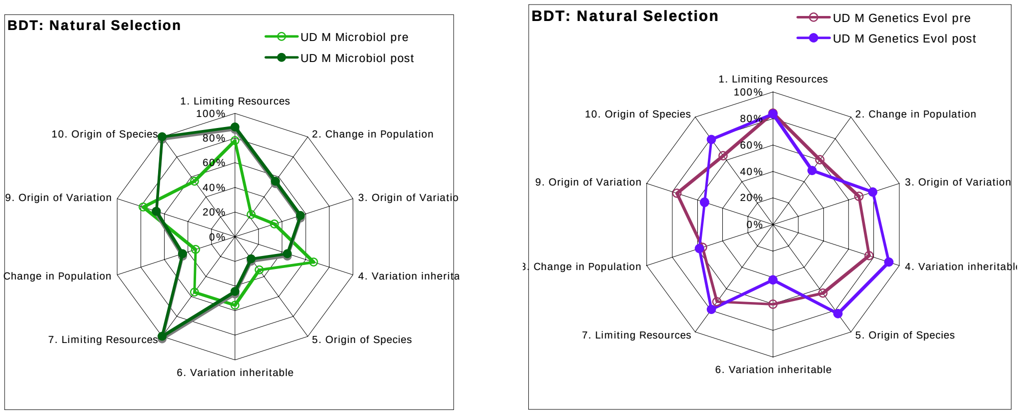
Figure 3.



In some classes we see changes occurring during the semester. Fig. 3 shows results from the upper division major's Genetics and Evolution course at the beginning and at the end of the course (pre and post). While students show the consistent difficulty with items 2 and 8, the pre and post scores are surprisingly high. We did see gains during the semester on some items (like 5 and 10 about origin of species) but the gains were not consistent on all items. This could be due to several issues. One is that there is no consistent sequence in which upper division students take these 3 courses. Since these results represent a set of students who voluntarily took this test, it is not a random sample, and we did not consider the course history of the students who took the survey

Natural Selection Pre- and Post-test Scores (Figure 4). Comparing pre- and post-test scores shows how the radar graphs can inform instructors about their learning in a semester. Students in Microbiology scored 100% on items 7 (carrying capacity) and 10 (origin of species), In contrast, Genetics and Evolution students showed gains on items 3 (origin of a trait), 4 (inheritance of genetic traits), and 5 (speciation).

Figure 4

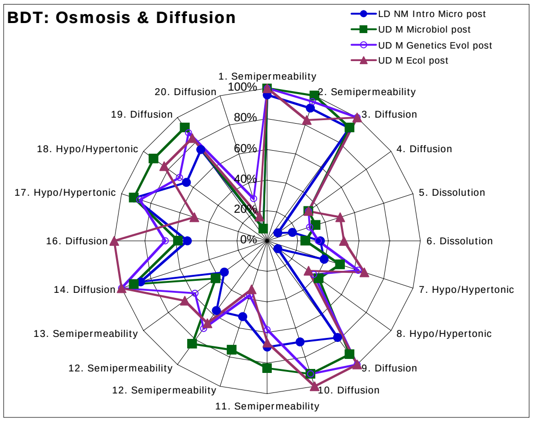


Because we did not account for which students were co-enrolled in other courses, it may be that students in Microbiology were concurrently enrolled in the Genetics and Evolution course. All participants were volunteers, not randomly sampled, so we cannot detect differences in incentives for participating in these results. However we do see some striking consistencies in some results (below).

Furthermore, following the constructivist idea, the outcome for instruction is still clear … explanations of the mechanisms of evolution, for example the source of variation and variation between members of a population, must give more treatment to why teleological and Lamarckian explanations are insufficient for explaining the unity and diversity of life. Students can then be guided to construct a more scientifically accurate understanding of the process of natural selection. We argue that with well developed diagnostic tests educators can identify "knowledge barriers" to address with effective instructional strategies in an appropriate curriculum sequence that will lead to increased knowledge gains on both diagnostic test items and more traditional items, such as on the Major Field Test (ETS). We hope to be able to make that comparison soon.

Osmosis and Diffusion Post-Course Scores (Figure 5). The osmosis and diffusion test has two-tiered items. Consider the ten diffusion questions. The first (odd-numbered) item asks a factual question, and the second (even-numbered) item asks for an explanation of why the response is correct. The power of the second items lies in the distracters that are chosen. Thus, students knew that molecules diffuse from high to low concentration in item 3, but many could not explain why this occurs in item 4. In items 5 and 6, students missed the fact that a small amount of salt will dissolve in a large container of water, as well as the reason why. They seem to know that heat speeds up diffusion and also why (9 & 10) and that blue dye will spread throughout the container and why (15 & 16). They know that concentration differentials affect diffusion (19), but much less sure about the reason why (20). There are clearly some concepts about diffusion that are not being learned or retained regardless of course. Others about semipermeability are apparently being learned by all. Comparing these results with those of a standardized test like the Major Field Test (ETS) will help confirm the acquisition or absence of understanding of those concepts.

Figure 3.



Discussion

We recognize that not all increases in mean class scores on diagnostic tests can be attributed to conceptual change in individual students. Some increases in average class performance will be produced by changes in the student population, as some students drop out of the program. We do not know which is the major contributor to increases in class scores, changing student populations or conceptual change in students. However, the desired end result of our undergraduate biology majors is to produce a group of students who have deep understandings of biology phenomena, consistent with what is known in the field of biology today. Both selective persistence in an undergraduate biology program and substantial intellectual effort on the part of each individual student are involved in producing this desired end result. Thus, we are using diagnostic tests to assess how successful we are in reaching this desired end point. We cannot claim that diagnostic tests are the best possible measurement tool for our purposes. What we can say is that we do not know of a better tool at this time.

Each student is tested at the beginning and end of each semester, with one form of one of the diagnostic tests. Thus, another possible concern is test fatigue and/or increasing test familiarity over time. This could contribute to increasing scores over time. On the other hand, students are often given points for completing a test, but their performance on a diagnostic test does not affect their grade. Thus, there is no particular motivation to either study for diagnostic tests or to cheat on them, or to try very hard to think about what the correct response may be. This could contribute to the lower scores seen on these tests as compared to traditional tests.

Diagnostic tests have, in come cases, revealed striking differences between classes of a single subject taught by different instructors and among different semesters. We are in the process of analyzing the background of those students, for example, to see if they took lower division preparatory science courses in smaller classes at community colleges or at SDSU.

Perhaps most importantly, diagnostic tests can give valuable feedback to individual instructors about what big biology ideas were learned successfully and which were not. Our challenge will be to generate interest among the faculty of our department to actually use the data we provide to modify their curricula and instruction. We are planning that approach and hope to report on those results soon.

Add more to discussion …

Relevant citations on diagnostic testing and assessing conceptual understanding in science

Anderson, D. 2003. Natural selection theory in non-majors biology: Instruction, assessment and conceptual difficulty. Dissertation for Ph.D. in Mathematics and Science Education submitted to San Diego State University and University of California – San Diego.

Anderson, D. L., Fisher, K. M., & Norman, G. J. 2002. Development and evaluation of the Conceptual Inventory of Natural Selection. Journal of Research in Science Teaching 39 (10): 952-978.

Bloom, B.S. (Ed.). 1956. Taxonomy of Educational Objectives: The classification of educational goals. Handbook 1. Cognitive domain. New York: McKay.

Christianson, R. G. & Fisher, K. M. 1999. Comparison of student learning about diffusion and osmosis in constructivist and traditional classrooms. International Journal of Science Education 21 (6): 687-698.

Clement, J. 1982. Students' preconceptions in introductory mechanics. American Journal of Physics 50 (1): 66 - 71.

DeVellis, R. F. 1991. Scale Development. Sage Publications, Newbury Park, CA.

Driver, R., & Oldham, V. 1986. A constuctivist approach to curriculum development in science. Studies in Science Education 13: 105-122.

Fisher, K. M. & Lipson, J. I. 1986. Twenty questions about student errors. Journal of Research in Science Teaching 23 (9): 783-803.

Gronlund, N. E. 1993. How to Make Achievement Tests and Assessments. 5th Ed. Boston: Allyn and Bacon.

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: 64-74.

Hildebrand, A. C. 1989. Pictorial representations and understanding genetics: An expert-novice study of meiosis knowledge. Ph.D. Dissertation, University of California - Berkeley.

Horn, J. L. 1965. A rationale and test for the number of factors in factor analysis. Psychometrika 30: 179-185.

Johnson, B., & Christensen, L. 2000. Educational Research: Quantitative and Qualitative Approaches. Boston: Allyn & Bacon.

Kaplan, R. M. & Saccuzzo, D. P. 1997. Psychological Testing: Principles, applications, and issues. 4th Ed. Pacific Grove, CA: Brooks/Cole Publishing Company.

Lautenschlager, G. J. 1989. A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. Mulitvariate Behavioral Research 24: 365-395.

National Research Council. 2000. Inquiry and the National Science Education Standards: A guide for teaching and learning (report). Washington, DC: National Academy Press.

National Research Council. 2003. Evaluating and Improving Undergraduate Teaching in Science, Technology, Engineering, and Mathematics. http://www.nap.edu/books/0309072778/html/

Novak, J. D., Mintzes, J. J., & Wandersee, J. H. 2000. Epilogue: On ways of assessing science understanding. In Mintzes, J. J. (Ed.), Assessing Science Understanding (pp. 355-374). New York: Academic Press.

Odom, A. L. & 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.

Palmer, D. H. 1999. Exploring the link between students' scientific and nonscientific conceptions. Science Education 83: 639-653.

Sadler, P. M. 1998. Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. Journal of Research in Science Teaching, 35: 265-296.

Sadler, P. M. 2000. The relevance of multiple-choice tests in assessing science understanding. In Mintzes, J. J. (Ed.), Assessing Science Understanding (pp. 249-278. New York: Academic Publishers.

Sundberg, M.D., and Dini, M.L. 1993. Science majors versus nonmajors: is there a difference? J. Coll. Sci. Teach. 23: 299 -304.

Sundberg, M D., Dini, M.L., & Li, E. 1994a. Decreasing course content improves student comprehension of science and attitudes towards science in freshman biology. Journal of Research in Science Teaching 31(6): 679-693.

Sundberg, M D. & Moncada, G.J.. 1994. Creating Effective Investigative Laboratories for Undergraduates. BioScience 44 (10): 698-704.

Sundberg, MD. 2002 Assessing student learning. Cell Biol Educ. 2002 Spring;1(1):11-5.

Sundberg, M.D. 2003. Strategies to Help Students Change Naive Alternative Conceptions about Evolution and Natural Selection. Reports of the National Center for Science Education 23(2): 23-26.

Tamir, P. 1971. An alternative approach to the construction of multiple choice test items. Journal of Biological Education, 5, 305-307.

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.