Using diagnostic test items to assess conceptual understanding of basic biology ideas:
A plan for programmatic assessment

Kathy S. Williams¹, Kathleen M. Fisher¹, Dianne L. Anderson², Mike U. Smith³, Jennifer E. Lineback¹

¹San Diego State University, ²Pt. Loma Nazarene University, ³Mercer University

Introduction

We must beware of “inert ideas” – that is, ideas that are merely received into the mind without being utilized or tested or tied into fresh combinations. That sentiment, voiced by Whitehead (1967), has been reiterated many times. Yet we know that large lecture classes, easy-to-grade tests, large numbers of students, and steadily increasing demands on faculty create conditions that promote students’ acquisition of “inert ideas.” An instructor has to work quite hard to follow a different path. Nonetheless, research overwhelmingly points to the need for students to be actively engaged in learning in order to develop deep understanding (e.g., Bransford et al. 2000, Crouch & Mazur 2001).

We know that people construct new knowledge by building on and connecting to what they already know. But unexpected consequences can occur. For example, researchers have discovered that most students who study photosynthesis, usually in middle and high school and again in college, fail to understand the basic idea! Why is that? Based on their everyday experiences, they assume that air has no weight. Logically speaking, if air has no weight, it cannot possibly be used to construct a massive tree (and people try to be logical).

Interestingly, these large-scale failures of comprehension remained undiscovered for many years. Instructors and books painstakingly explained photosynthesis to students. Students performed adequately on simple fact-recall test items. How could they fail to understand? What happened was that, as we now know, the instruction ignored the essential starting point: the student’s prior understanding. If an individual is certain that air has no weight, it is unlikely that an instructor is going to convince the individual that it can be used to construct a tree. Instruction was trumped by what we call the students’ “prior knowledge” or “commonsense ideas.” The idea that naïve conceptions play an important role in students’ understanding was first discovered in the late seventies and early eighties (e.g., Clement 1982, McCloskey 1983). Naïve conceptions or misconceptions, as they were commonly called, are characterized as ideas that a) tend to be shared by a significant proportion of a population, b) produce consistent error patterns, and c) are remarkably resistant to being “taught away” (Fisher & Lipson 1986). Lecture teaching and other forms of telling have been observed to be particularly ineffective (Bransford et al. 2000). Today, researchers are learning to use students’ commonsense ideas as stepping stones upon which to build more advanced understanding (Smith et al. 1993). Unfortunately, these giant steps have yet to influence the daily teaching events that occur in millions of classrooms across the country.

How do you convince the academy of such “truths” that a) knowing about students’ commonsense ideas matters, b) active learning requires active engagement, and c) change is urgently needed? University faculty have been lecturing for centuries. Tradition is revered. There is little institutional reward for innovative teaching. The physical structures called classrooms are
designed to support lecturing. And faculty are very busy. Thus, introducing change is problematic. This is where diagnostic tests may be able to make a contribution.

The Force Concept Inventory (“FCI;” Hestenes & Wells 1992, Hestenes et al. 1992, Halloun et al. 1995) is a diagnostic test that provides a forced choice between Newtonian concepts and commonsense alternatives. It has been used in hundreds of high school and college physics classes. Physics (unlike biology) has a distinct advantage in having a common curriculum across introductory course offerings. Hake (1998) produced a meta-analysis of 62 physics courses enrolling 6,542 students at high school, college and university levels. There were 14 traditional courses \( (n = 2,084 \text{ students}) \) and 48 interactive engagement courses \( (n = 4,458) \). Hake used normalized gain scores obtained with the FCI \[ \text{ratio of the average actual gain (\% post – \% pre)} \text{to the maximum possible average gain (100 – \% pre)} \] to obtain consistent analyses over diverse student populations. He found that the 14 traditional courses had an average gain of 0.23 +/- 0.04 (standard deviation), while the 48 interactive classes were almost two standard deviations higher with an average gain of 0.48 +/- 0.14.

Similar findings emerged from a meta-analysis of STEM (Science, Technology, Engineering, and Math) courses that engage students in small-group collaborative learning (Springer et al. 1980). The authors analyzed 39 of 383 studies that met the authors’ strict criteria for inclusion (http://www.wcer.wisc.edu/archive/cl1/CL/resource/scismet.htm). They found that students who learn in small collaborative groups generally demonstrate greater academic achievement, express more favorable attitudes toward learning, and are more successful in persisting in STEM courses or programs. As above, the differences between students engaged in active learning and those in traditionally passive classes were large.

**Programmatic assessment**

Programmatic assessment aims to assess the effectiveness of a curriculum. Colleges and departments across the country are currently engaged in trying to develop effective means for conducting such assessments. When you Google programmatic assessment, you can see the plans developed at many institutions. Professional programs are generally in the lead. SDSU’s nursing program, for example, follows students post-graduation and asks their employers to evaluate the performance of SDSU graduates and make suggestions for curriculum changes. This could be a costly effort for most undergraduate programs.

Programmatic assessment begins with setting goals and objectives. What do we aim to accomplish? CSU has developed a general set of goals and objectives for all CSU graduates. The SDSU biology department has created a more detailed list for biology graduates. In this context, we are developing and administering diagnostic tests each semester to assess gains our undergraduates make in deep understanding of important ideas.

Our goals are twofold. We aim to a) develop a series of diagnostic tests suitable for biology and b) implement those tests as programmatic assessment tools to help faculty make informed choices about how to improve the various biology curricula.

**Methods**

We are using diagnostic tests in biology, administered to subsets of students pre- and post-semester, to assess knowledge gains across our biology majors’ curriculum. The five diagnostic tests currently in use focus on either “big” ideas (natural selection, energy transformation, and nature of biology) or foundational principles necessary for understanding many higher level concepts (osmosis & diffusion, mitosis & meiosis).
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 about processes in science. Since the alternative conceptions used as distracters 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.

**Test Item Styles**

We have used four styles of test items. In *two-tiered items*, 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.

*Situational items* present an authentic scenario in a short paragraph and pose several questions about organisms in the context of the scenario. As with tiered items, each response option is aligned with either the scientifically correct or alternative conceptions.

*Simple multiple choice items* consist of a header with four or five response options similar to those described above.

*Likert scale true-false items* provide choices of Agree-Disagree or the more extended responses of Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree.

**Diagnostic Tests**

If we wanted to study the process of conceptual development, 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. The correct scientific response is what we hope our biology majors will be able to recognize and select, while the alternative conceptions being used as foils reflect scientifically incorrect reasoning that we hope our students will reject. To repeat, because this is easily misunderstood, we are not studying students’ prior knowledge or cataloging common misconceptions. We’ve simply chosen effective foils to aid in assessing science knowledge.

1. **Natural Selection Diagnostic Test.** Items were selected from the Conceptual Inventory of Natural Selection (“CINS,” Anderson et al. 2002) and modified in many cases, with permission from the authors. Included in the natural selection diagnostic test are two *situational item* clusters about scientific studies of the Galapagos finches and the Venezuelan guppies, with five questions each.

2. **Osmosis & Diffusion Diagnostic Test.** Items were selected from the Osmosis and Diffusion test of Odoms & Barrow (1995), and modified as deemed appropriate with permission from the authors. This diagnostic test consists of ten *two-tiered items*, twenty questions in all.

Tests (1) and (2) are now stable. By “stable,” we mean that we are no longer modifying items and the test meets the following criteria: a) all incorrect responses are selected by some students, b) the proportion of students selecting the correct response falls within the range of 30% to 70% as recommended by Kaplan & Saccuzzo (1997), and c) items have been validated with student interviews.

Tests 3-5 are currently in various stages of development and are not yet available in stable form.
3. **Mitosis and Meiosis Diagnostic Test.** This test consists of *two-tiered items* and is in late-stage evaluation and revision.

4. **Energy & Matter Transformation Diagnostic Test.** This test consists largely of *two-tiered items* and some simple multiple choice items. It focuses on photosynthesis, decomposition, and energy and matter transformations.

5. **Nature of Biology Diagnostic Test.** We developed a set of question stems drawn largely from (Abd-El-Khalick & Lederman) and initially administered the test with Likert-scale Agree/Disagree responses. We are currently developing a revised form with simple multiple choice responses that will be piloted next semester.

**Example Test Items**
[* indicates correct responses]

Below is a situational item from the *Natural Selection Diagnostic Test.* The header concerning Venezuelan guppies is derived from a scientific study and is tied to five test items.

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.*

This is a *two-tiered item* from the *Osmosis and Diffusion Diagnostic Test.*

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 that
   a. there is movement of particles from a high to low concentration. *
   b. the salt 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.

And this is a *two-tiered item* from the *Mitosis and Meiosis Test.*

1. A human egg and sperm join together to form the beginnings of a new baby and an embryo begins to form. As 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. only responses (a) and (b) are correct.
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.

Two versions of Likert scale items have been given in succession with the Nature of Science Test:
23. The ultimate test of a scientific model is its ability to accurately predict outcomes.
   Agree___ Disagree
2. Any scientific idea or assumption can be challenged by contradictory evidence.
   Strongly Agree___ Agree___ Neither Agree nor Disagree___ Disagree___ Strongly Disagree___

We are currently shifting from a “Nature of Science” test to a more focused “Nature of Biology” test, and from Likert-scale responses to the simple multiple choice format, as shown below.
13. Theories in biology, like Natural Selection, represent established knowledge and thus will not change in the future.
a. True
b. False*
14. The reason for my response above is because
a. biology knowledge is tentative and subject to revision if warranted by new data.*
b. established knowledge has been proven and is the truth.
c. theories are hypothetical and never represent established knowledge.
d. laws are the only form of biology knowledge that are fixed and unchanging.

Test Development
We began test development by identifying known misconceptions (including both published findings and observations collected by the authors). Next, particular topics were selected for inclusion in the tests. In two cases we began with existing diagnostic tests, as noted previously. A specification table was created for each test summarizing the topics of interest, the common alternative conceptions for each topic that are used as distracters, and the correct or scientific response to the header (Table 1). Some tests contain two items per topic, typically presented with different background settings. For example in the Natural Selection Test, an item about limited resources appears in both the finch and guppy situations). The tests are evaluated and refined, usually over the course of 2-4 semesters. We have used the following guidelines for test item creation:

1) Eliminate jargon. 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 for genetics concepts, since there really are no commonly used terms for such ideas as gene, allele, and DNA.

2) The correct response should be clear and unambiguous.

3) Each distracter (incorrect response) should reflect a common naïve conception and should be attractive to at least some students.

4) Scientific data in test items should be drawn from an actual scientific study. We feel this is important to keep our representations of scientific research “authentic.”
5) Students should be interviewed to determine whether they interpret the test headers, stems, and responses in the intended manner. We can detect unanticipated misinterpretations by asking the interviewees to restate the headers and responses in their own words.

6) At least two experts should be interviewed who have not been involved in test development complete the tests to determine if all agree that there is only one correct response.

Table 1. Specification table for the Natural Selection Diagnostic Test. Shown in parentheses are the corresponding item numbers and response letters.

<table>
<thead>
<tr>
<th>Topic/Issue</th>
<th>Common Confusion or Naïve Conception (subtest item/stem #, option A-D)</th>
<th>Scientific Ideas (subtest item/stem #, option A-D)</th>
</tr>
</thead>
</table>
| A - Population stability  | • All populations grow in size over time (1A, 7B)  
• Populations decrease (1D, 7C)  
• Populations always fluctuate widely or 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)                                       |
Biology Programs Included in Programmatic Assessment

The Biology Department offers a Biology B.S. with emphases in Cellular and Molecular Biology, Ecology, Evolution and Systematics, Marine Biology, and Zoology. Students may also earn a B.A. in Biology, or a B.S. or B.A. in Microbiology. All programs have identical lower division science course requirements, and all require the same three “core” upper division courses in a) ecology, b) genetics & evolution, and c) 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 3 upper division core courses and the upper division microbiology course. As noted previously, the diagnostic test items focus on biology ideas that are either foundational (e.g., osmosis & diffusion) or big (e.g., natural selection), meaning they are topics that serve as either foundational knowledge or represent over-arching frameworks that influence learning of many ideas in biology. These topics are generally not taught directly in the particular courses in which students are being tested, but instructors are likely to assume that students have the knowledge being tested. 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.

Diagnostic Test Administration

As part of the development and validation process, and as first steps in programmatic assessment, the diagnostic tests are administered as pre- and post-semester tests in biology courses for both biology majors and non-majors at our large public university. These include lower division non-majors’ and upper division majors’ core courses. 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 receives a post-test on the same topic.

Test administration methods have been continuously evolving. We began by administering the tests in class, which requires using valuable class and personnel time. We then moved to on-line administration and have explored several different technological systems in succession. We are currently administering the tests via the Blackboard class management system, now that it provides individual item responses that we need for analyses.

Students are given credit (a few points) for completing one pre- and one post- test per semester. The pre-test is given on-line during a 10-day window in the third week of the semester, after class enrollments are finalized. The post-test is given during a 10-day window spanning the last weeks of classes. Survey availability is announced in the classes and via Blackboard, and students gain access to the tests through their Blackboard accounts. After the window closes, the course instructors are provided with names of students who completed the tests, so they can award credit.

To date, each biology faculty member has made his or her own decision about how much credit to give the students in their class for their completion of the tests. However, we would prefer to achieve consistency across courses, and we are also concerned about the possibility of students responding to tests haphazardly in the interest of gaining a few points with the least effort possible. As a consequence, we are suggesting to department members that all students receive 3 points if they score 50% or more on a diagnostic test, 2 points if they score 25-49%, and 1 point if they score 0-24% (that is, a point simply for taking the test). These points would apply to both the pre- and post-semester tests.
Awarding variable points will likely discourage flippant test-taking, but does not address the possibility of cheating. Students take the tests on their own time and at their own locations, and we hope that the low number of points provides little incentive for cheating. If cheating is extensive, it could cause over-estimation of our students’ abilities. However, due to the nature of the diagnostic test items, we think it unlikely that students easily could look up correct answers.

In addition, we are tracking students so when a student is enrolled simultaneously in two or more target courses, they are asked to complete only one pre- and post-test. Some effort is also made to try and assure that a student receives a different test each semester. With hundreds of students, the tracking and record-keeping is quite challenging.

Evaluation and Analysis

*Discriminability* (point biserial values) and *difficulty* values are calculated 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 tests designed to identify concepts that students do or do not understand, rather than to discriminate among students, the point biserial values are of limited usefulness here, compared to their use on norm-referenced tests (Gronlund 1993).

The *difficulty* is determined by the proportion of students who respond to the item correctly. These values are used to create radar graphs that provide visual descriptions of performance on each test, by item, among classes, and between pre- and post-tests. We also are now recording the time it takes each student to complete the tests, to relate that to performance.

Our ultimate goal is to analyze results using covariates indicating such things as which courses students previously completed, the campus where lower division preparation was completed, and educational methods used in the prior courses as well as in the “current” course in which the test is being administered.

Results

Our results over the past three semesters indicate several patterns and suggest additional areas of research and reform in our curriculum. In some cases the patterns are consistently similar across semesters and courses. However, we also see differences among courses and semesters that are prompting more research.

*Some misconceptions persist across semesters and courses.* The most obvious result is the striking consistency in the scores on the various items in the tests. Results from the *Natural Selection test* illustrate this (Fig. 1). The radar graph displays the proportion of students, by class, that selected the correct response on each of the 10 *situational items* on the test. Each ray of the graph represents a question, and the proportion answering correctly is represented by the concentric circles, ranging from zero% in the center to 100% on the outer circle. Concepts tested are indicated beside the item numbers.

Students at the end of most classes in all three semesters found items 2 and 8 to be especially difficult. Those two items ask students to recognize that populations evolve as the frequencies of different traits in a population change over time, with item 2 asking about finches and item 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 (see Table 1 for alternative conceptions). Rather, they think that individuals change. Many
students also found items 5 and 10, both asking about how species evolve, to be challenging as well. These patterns persisted across semesters and across courses, though we do see deeper understanding in upper division Biology majors compared to lower division non-biology majors, as indicated by position of the non-major’s blue line within the performance lines for the upper division Biology majors (Fig 1).

This demonstrates the striking consistency with which students are attracted to particular non-scientific conceptions (or commonsense ideas) about some of these foundational concepts, even after what many faculty consider advanced coursework. 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.

Figure 1. Natural Selection test post-semester results. Sample sizes range from 6-34 students per class, per test. The radar graph displays the proportion of students, by class, that selected the correct response on each of the 10 items on the test. Each ray of the graph represents a question, and the proportion answering correctly is represented by the concentric circles, ranging from zero% in the center to 100% on the outer circle. Concepts tested are indicated beside the item numbers (see Table 1 for details). (LD=Lower Division, UD=Upper Division, NM=Non-Majors, M=Majors)
Students get the “how” or “what” of an idea – but not “why.” A second consistent pattern seen in the data is exemplified by scores on the two-tiered items as in the Osmosis and Diffusion Test (Fig. 2). 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 results of these items consistently show us that students may be able to correctly answer the more factual “what happens” first item, but are not able to select the correct “why” or explanation from the alternatives offered in the second item.

For example, students generally seem to know that molecules diffuse from high to low concentration in item 3 (Fig. 2), but many could not explain why this occurs in item 4. They know that concentration differentials affect diffusion (item 19), but are much less sure about the reason why (item 20). This produces the “zig-zag” pattern seen in these graphs and also the results from the tiered items in the Meiosis and Mitosis Test.

Figure 2. Two-tiered items on Osmosis and Diffusion Test, showing the ability of students to recognize what might happen but not how the process works. 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. Sample sizes range from 7-36 students per class, per test. (LD=Lower Division, UD=Upper Division, NM=Non-Majors, M=Majors)
Knowledge increases with experience – sometimes. There are clearly some concepts about natural selection and diffusion that are not being learned or retained regardless of course, while others are apparently being learned by all. In some results, like those from the Natural Selection Test, we see knowledge gains with experience in the study area. For example, as we might predict, the non-biology majors’ performance (blue line) is consistently inside the performance lines for the upper division Biology majors, indicating lower scores for more novice learners (Fig 1). We might expect that at the end of the Genetics course, upper division biology students would perform substantially better than other students, and sometimes they did (items 4 & 5 in Fall 06). However, this was not always the case, with students in other upper division biology courses often performing equally well or better (items 2 & 8, and 9 & 10). This may be due to the fact that there is no standard sequence in which upper division students take these 3 upper division courses. So it may be that many of the students in the Microbiology and Ecology courses had already taken the Genetics course and retained the knowledge (or not). We are now examining whether course sequencing affects performance on these tests.

On other tests we see consistently low performance on some items that doesn’t change over the semester or across classes. This is seen in the Osmosis and Diffusion Test results for items 3 through 6 and 19& 20 (Fig. 2) and has been confirmed with student interviews. Across all three semesters, the non-majors post-semester performance is within the performance range of the upper division biology majors. This was seen in results from items 3 & 4 and items 19 & 20 about diffusion of particles (mentioned earlier), and in results from items 5 & 6 where students missed the fact that a small amount of salt will dissolve in a large container of water as well as the reason why. These results indicate that if these are concepts that are prerequisites for the courses, then students are not learning or retaining the concepts before they enter these courses - and we need to redesign the curriculum to improve that situation. However, if the ideas are meant to be taught in the courses where the testing occurs, large proportions of students aren’t getting it and instruction should address that.

Alternatively, the good news is that we consistently see high performance on items that doesn’t change over the semester or across classes, or even within semesters, since in these cases pre-and post- semester scores are equally good. This is seen in the Osmosis and Diffusion Test results (Fig. 2). For example, all students seem to know that heat speeds up diffusion and also why (items 9 & 10), and that blue dye will spread throughout a container of water and why (items 15 & 16). These consistently high scores indicate two things to us. First we want to know when the students are learning this information, and we plan to begin testing in the lower division courses for biology majors next semester. We may also test in the non-majors general biology course that precedes the introductory microbiology course for non-majors. It may be that the students are learning and retaining this knowledge in the introductory courses, indicating good course design that we want to propagate. If most students are entering our introductory courses with that accurate prior knowledge, then we will consider dropping those questions or modifying them to be more informative about our students’ learning.

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 program is to produce undergraduate biology majors 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 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.

Perhaps most importantly, diagnostic tests can give valuable feedback to individual instructors about what big and foundational 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.

Steady gains in understanding with increasing years of study of biology can be seen in the case of natural selection. This is reassuring, since evolution and natural selection provide the overarching theory of biology. It no doubt reflects in part the increased attention being given to making evolution comprehensible in both texts and lectures as well as in public programs. What is disappointing, however, is the lack of any measurable progress in other areas. Is this attributable to poor selection of topics, or poorly designed tests, or actual absence of deeper understanding by the students?

We consider our selection of topics to span a breadth of foundational concepts basic to most biology knowledge. Understanding osmosis and diffusion seems critical for understanding such areas as molecular biology, microbiology and physiology, since these processes represent the primary forces affecting the movement of molecules in living things. It is also important in ecology because of their roles in plant life and in movement of materials in nutrient cycles and through the soil. Likewise, understanding mitosis and meiosis seems essential because of their critical roles in such areas as genetics, natural selection, artificial selection, and innate disease factors. We suspect students’ understanding of these topics holds steady or declines as they progress through the major because professors assume that this is something students already know, and don’t address the topic and attempt to dispel naïve conceptions.

Students’ lack of understanding of the flow of energy and matter is dramatic. As many as 90% of biology students may fail to recognize that plants construct themselves from carbon dioxide in the air, in large part because they believe that air has no weight. This represents an alarming lack of understanding of the periodic table and its implications. The cycles of matter are typically presented in the backs of biology texts and are rapidly glossed over. Yet there are tenacious naïve conceptions interfering with comprehension of these cycles, and the subject needs to be actively addressed.

Similarly, inadequate understanding of the nature of science by the general public and by science students has been well documented (Lederman 1992, McComas 1997). We anticipate finding a similar lack among biology students in understanding the nature of the science of biology.

Thus, we conclude that the absence of a trend toward deeper understanding of osmosis and diffusion, mitosis and meiosis, and the flow of energy and matter as students progress through the biology major is real. Our hope is that the results of the tests will be informative to faculty and will prompt changes in the curriculum, including both content and perhaps also instructional methodology. We recognize though, that our own research with diagnostic tests may not drive the latter until there are alternative instructional methodologies to compare.
References


