Two-dimensional, implicit confidence tests as a tool for recognizing student misconceptions

Linda B. Taylor*, Shana R. Spindler*, R. Kathy Garvin-Doxas & Mike Klymkowsky+

Bioliteracy Group, Molecular, Cellular & Developmental Biology
University of Colorado, Boulder
Boulder, Colorado 80309-0347 USA

* equal contribution
+to whom correspondence should be addressed
T: 303-492-8508 | F: 303-492-7744 | E: klym@spot.colorado.edu W: http://bioliteracy.net

Journal of College Science Teaching - in press.

Abstract: The misconceptions that students bring with them, or that arise during instruction, are a critical barrier to learning. Implicit-confidence tests, a simple modification of the multiple-choice test, can be used as a strategy for recognizing student misconceptions. An important issue, however, is whether such tests are gender-neutral. We analyzed the results of exams administered to students (both majors and non-majors) in MCDB 1111: Biofundamentals at the University of Colorado at Boulder. At a statistically significant level (> 95%), there was no difference between women and men regardless of whether their answers were confidently correct or incorrect, suggesting that such two-dimensional tests are a gender-neutral tool.

Introduction: It is well established that a major barrier to learning are the misconceptions that students bring with them into the classroom or that develop during the course of instruction (see Committee on Undergraduate Science Education report, 1997). Conventional teaching strategies, even those that directly address identified misconceptions, often fail to change students' deeply held ideas (see as an example Gregg *et al,* 2001 which describes a study on students' views on vision and the effects of various interventions). While critical to recognize and address, the exact nature of student misconceptions can be difficult to identify through standard classroom interactions, particularly in larger classes. In a study on the prevalence of misconceptions about blood circulation and gas exchange among prospective elementary teachers, Paleaz *et al,* (2005) noted that different ways of asking students to display their understanding revealed distinct types of misconceptions. For example, student drawing appears to be a particularly effective strategy in uncovering misconceptions. A similar insight is illustrated in the film "*A Private Universe*" (1989) with respect to the relative movements of the earth, moon, and sun.

Generally student misconceptions persist until the student recognizes that their "understanding" does not work and that following it leads to incorrect answers or logical conclusions. For a question or task to lead a student to recognize and address their mistaken assumptions, it must force the student to actively use their conceptual understanding. A particularly effective strategy to lead students to such a "eureka" moment involves the use of tutorials, conceptual problems that require students to apply information and ideas they have learned to solving novel problems. Tutorials have been developed in physics (see Redish, 2003 and references therein) and astronomy (e.g. Adams *et al,* 2005), and we have begun to develop, test, and revise similar tutorials for introductory biology (http://www.colorado.edu/MCDB/MCDB1111/Tutorial%20TOC.htm).

Without identifying and addressing the misconceptions students harbor, robust and resilient understanding is nearly impossible to achieve. Instructors commonly assume that conventional testing strategies, whether multiple choice or essay, evaluate students' conceptual understanding, but as Pelaez *et al,* (2005) show, this is not the case. More often than not, testing instruments require that students simply recognize or remember the correct answer rather than use their understanding to construct it. In this context, there is little impetus for real learning; misconceptions, if present, remain firmly in place. Exams of this type, whether multiple choice or essay, have been labeled "inauthentic" by McClymer & Knoles (1992). Good grades on such tests often lead both students and instructors to the mistaken belief that learning has occurred, when in fact students may still harbor significant misconceptions that actively interfere with true understanding. A critical ramification of such "ersatz learning" is that teachers from the elementary through the college level can transmit their own unresolved misconceptions to their students (see http://www.ems.psu.edu/~fraser/BadScience.html). Given the high percentage of high school teachers, as well as members of the general public, who favor teaching various forms of creationism, often to the exclusion of biology (Scott, 1999; Pew Research Center Report, 2005), it is probably no exaggeration to suggest that both high school and college biology courses are failing to address serious misconceptions about biology in general and evolutionary biology in particular.[1]

One attempt to identify and address student misconceptions is the use of Concept Inventories. The first, and most influential of these is the Force Concept Inventory (FCI)(Hestenes *et al,* 1992). Designed to probe conceptual understanding of Newton's laws of motion, rather than the ability to memorize terms or manipulate equations, the use of the FCI revealed that high student grades often did not correlate with a robust conceptual understanding (Hake, 1998; Powell, 2003). Since the introduction of the FCI, other concept inventories have begun to appear. Among these are the Force and Motion Concept Evaluation (FMCE)(Sokoloff & Thorton, 1998), the Brief Electricity and Magnetism Assessment (BEMA)(Ding et al., 2006), the Quantum Mechanics Concept Survey (McKagan & Weiman 2005), the Natural Selection Concept Inventory (Anderson et al., 2002), and our own Biology Concept Inventory (BCI)(available in August 2006 through bioliteracy.net). However, because they are generated through research to uncover student misconceptions and validated through student interviews, concept inventories are both expensive and time-consuming to generate. More importantly, they are not meant to be used for the assessment of individual students, but rather as a pre-/post-instructional instrument to evaluate learning gains associated with specific teaching strategies.

[1] http://bioliteracy.net/Comment-Alberts.htm

Given their role as a teaching assessment tool, concept inventories do not address an instructor's immediate and ongoing need to identify their own students' misconceptions. This raises the practical question: are there simple and efficient ways for instructors to recognize at least some of their students' misconceptions in time to address them? Clearly, the most direct method, engaging in Socratic interactions during which students are encouraged to explicitly define their terms and enunciate their assumptions and approach to a particular problem, is generally impractical. An increasingly popular alternative is the use of electronic student response systems (Ward, 2003; Wood, 2004; Dye, 2005), by which instructors can replace the generally rhetorical, "Are there any questions?" and, "Does everyone get it?" with an immediate, and often enlightening, assessment of student understanding. A complementary approach involves a simple modification of the standard multiple-choice test.

Information Reference or Two-Dimensional Multiple Choice Testing: Two-dimensional tests (TDTs) aim to determine students' certainty, or confidence, in their answers (Bruno, 1993). Certainty in answers does not necessarily translate into student self-confidence in their abilities, as is usually defined in the literature. However, the confidence nomenclature is well established in the context of TDTs and we will follow this convention here. There are two versions of a confidence-based TDT: self-assessed, or explicit and implicit (**FIG. 1**). In a self-assessed, confidence-based TDT, each question consists of two parts: a conventional multiple-choice question paired with a second part that asks the student to indicate their level of confidence in their answer. For example, they can indicate whether they are "very confident," "semi-confident," or "guessing." In an implicit confidence test, the level of student confidence is reflected by a single response. Student are assured that one and only one response is correct, but they are given the option to "waffle'" - if they are not sure whether the correct answer is A or B, they can pick "A or B". If they have no idea as to which is the correct choice, they can indicate that as well. While both types of TDTs assess student confidence, implicit tests have the advantage in that students do not have to explicitly assess their confidence level, their choice of answer contains that information.

Self-assessed

The common ancestor of all life on earth…

- O is estimated to have lived ~3 billion years ago
- O is likely to arrived on earth from outer space
- O had a morphology very much like a eukaryotic cell.

I am
- O very confident
- O semiconfident
- O guessing

Implicit

Pick the true statement:

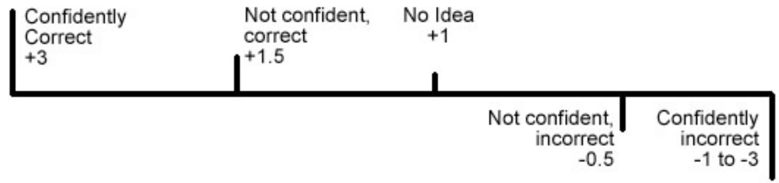
A. Based on the morphology of an organism it is possible to determine unambiguously whether it is primitive or degenerate
B. Over time the process of evolution will always generate more and more complex organisms.
C. Morphology can be deceptive, a primitive looking organism may in fact be descended from a more complex predecessor.

D. A or B   E. A or C   F. B or C   G. No Idea

Figure 1: Examples of explicit (top) and implicit (bottom) confidence test questions.

A critical component in both types of TDTs is the way that points are assigned to particular response (**FIG. 2**). Students receive maximum credit for confidently correct answers and the maximum penalty, in the form of negative points, for confidently incorrect answers. As students reveal their uncertainty, the number of points they receive or lose is reduced. One consequence of this grading strategy is that, since an incorrect answer is penalized more than admitting one's ignorance, the students' optimal test-taking strategy is dramatically altered (Bruno, 1993). For this reason, the structure of the exam may initially upset many students until they adapt to it - adaptation usually occurs quite quickly.

Figure 2: Grading scheme for two-dimensional test questions.



While there are important benefits associated with changing the strategies students use in answering questions, the pedagogical value of using TDTs involves formative assessment – it does not take much time or effort to identify those questions where a large percentage of students have answered confidently wrong. These questions serve as signposts for student misconceptions; the more closely the incorrect answers (distracters) reflect student misconceptions, as they are designed to do in a well-researched and validated concept inventory, the more student responses mirror their actual misconceptions. Even with poorly designed distracters, however, TDT responses provide quick and often valuable feedback to the instructor. Assuming that the questions themselves are not seriously flawed, if the majority of students taking the test are confidently correct on a particular question, then it is likely that the instructor has effectively conveyed the relevant concepts. If, however, the majority of students are confidently wrong on a particular question, it is likely that the students are misinformed or that the instructor has failed to lead the students to a robust understanding, perhaps because of interfering (and unrecognized) student misconceptions. If many students answer "semi-confident" or "guessing/no idea," the instructor may need to evaluate their approach to presenting the materials or encouraging student engagement with it.

To use the information garnered from TDTs in learning interventions, it is important that students have an incentive to revisit and correct their understanding. A number of strategies are possible. In MCDB 1111:Biofundamentals we use what we call "I know it now" exams. These exams are in the implicit TDT format, but without the "don't know/guessing" option. IKIN exams are given during the final exam, and the points gained (or possibly lost) are added to the original midterm exam scores. In this scenario, the original exam score is seen as provisional and can be improved through further study.

Confidence-Based Testing and Gender Concerns: Beginning in the spring 2003 semester, we implemented confidence-based TDTs in MCDB 1111:Biofundamentals. During the spring of 2004, we conducted an informal survey (**FIG. 3**) to assess student attitudes to multiple-choice tests (42 students responding out of 75). Approximately 42% of the responding students liked multiple-choice exams because they could guess and use partial knowledge to get the answer "right," and/or they did not feel that they had to study as much for multiple-choice exams. That said, 51% indicated that simple multiple-choice exams did not assess their true knowledge of the subject. A number of students had concerns that confidence testing would lead to a bias against females, on the assumption that females would be less likely to mark "confident" on explicit confidence tests and so would be less likely to score full points. The literature generally echoes the student's concerns; it is generally accepted that female students in science disciplines are less confident than males (e.g. Seymour & Hewitt, 1997; Light *et al*, 2000). That said, there is a significant difference between self-confidence and certainty in test answers. To address this concern and determine whether female students were less likely to answer questions confidently, we performed a statistical analysis of implicit (embeddedconfidence) TDTs to determine (i) whether females answer semi-confident rather than confident more often than males, and (ii) whether males score better on average than females when implicit TDT's were used.

Figure 3: MCDB1111 Multiple-choice Test Survey

We are conducting a survey of student opinions about testing. We would appreciate your responses to these questions

1. Please finish the following statement. I like multiple-choice questions because ....

- I can use partial knowledge to select answers, and no one will know that I am not confident in my answer.
- I do not have to study as hard or commit as much to memory because I only need to recognize the answer
- Without knowing anything at all about the question, there is a still the chance that I can get it right
- None of the above capture my feelings about multiple choice questions
- I do not like multiple choice questions

2. If you do not like multiple choice questions, tell us why.

3. Do you think that traditional multiple-choice questions accurately evaluate your knowledge of the subject matter? **yes or no**

4. What do you think is the best way to evaluate student learning?

Statistical methods: The MCDB 1111:Biofundamentals course (Spring 2003/2005) at University of Colorado, Boulder was used for all data collection. The 2003 class consisted of 39 females and 36 males, and the 2005 class consisted of 25 females and 26 males; approximately 30% of all students in both years were MCDB majors. All data was collected after the drop/add deadline for the course. In 2003, three mid-term TDTs were analyzed. In 2005, the final TDT (but not the IKIN exams) was analyzed. Each midterm exam had 20 to 25 implicit TDT questions. Two-tailed t-test analysis (>95% confidence level)(Harris, 1998) was used to identify any significant difference between the number of times females and males chose confident versus semi-confident answers. In addition, two-tailed t-tests were performed to determine if there was a significant difference in overall exam scores between males and females. We also reexamined the data using the less rigorous Chi square test (see supplemental figure 1) and again found no significant difference between the numbers of confidently answered questions (whether correct or incorrect), between males and females. .

For all three exams in 2003, males and females did equally well (**FIG. 4 upper panel**), and there was no statistically significant difference in the frequency in “confident/semi-confident/guess” choices between the two groups. Out of a total of 3601 "confident" male answers and 3960 "confident" female answers, males were correct 82% of the time, while females were correct 80% of the time. Similar results were found in 2005 (**FIG. 4 lower panel**), with males and females displaying similar levels of confidently correct and confidently incorrect responses.

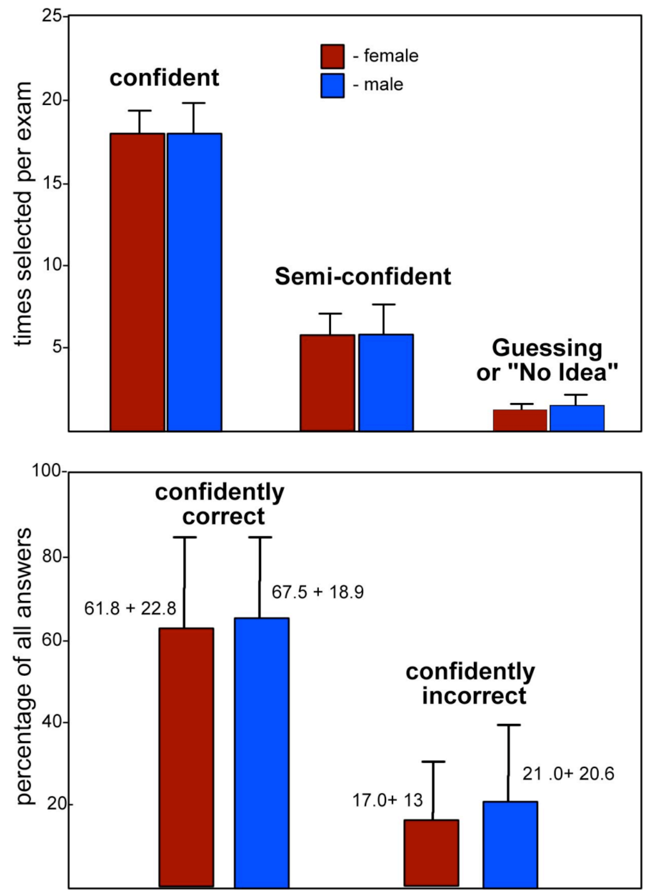


Figure 4

Summary: Even the most rigorously designed simple multiple-choice tests consistently overestimate student learning and provide little feedback to the instructor about the nature or source of student mistakes (Bruno, 1993). If a student guesses on an exam question and does so correctly, the student rarely revisits the information. Misconceptions, if present, go unrecognized and unaddressed. These misconceptions are not benign; they can lead to a cascade of misunderstandings that result in intellectual confusion and an inability to build on and apply past knowledge to new situations. In contrast, implicit confidence TDTs change the testing landscape and provide instructors with valuable feedback about student confusions, as well as identifying material that the students are uncertain about or areas in which they are actively misinformed. If linked to other student response systems and targeted interventions (e.g. tutorials), implicit confidence TDTs can become an important tool in increasing student understanding and serve as a "reality check" on runaway course content, which we refer to as syllabus bloat (Klymkowsky, 2005). It is trivial to transform standard multiple choice tests into implicit TDTs, and straightforward to identify questions that provoke confidently incorrect or confused student responses. Moreover, it appears that implicit confidence TDTs are not gender biased and so may be implemented without placing females at an instructional disadvantage.

Acknowledgments: We thank Jason Spindler for advice on editing the manuscript, Katherine Tepe for help with the statistical analyses, Becky and Sara Klymkowsky for editing and instructional insights, and the students of MCDB 1111 2003-2005 for their input in the study.

Literature cited:

Adams, J., Prather, E.E., Slater, T., & Dostal, J. (2005). Lecture Tutorials for Introductory Astronomy. Prentice-Hall.

Anderson, D.L., Fisher, K.M., & Norman, G.J., (2002). Development and validation of the conceptual inventory of natural selection: Journal of Research in Science Teaching, **39**: 952-978.

Harvard-Smithsonian Center for Astrophysics (1989). A Private Universe. Available on line at http://www.learner.org/resources/series28.html

Bruno, J. E. (1993). Using Testing to Provide Feedback to Support Instruction: A Reexamination of the Role of Assessment in Educational Organizations. In "Item Banking: Interactive Testing and Self-Assessment, D. A. Leclercq and J. E. Bruno, editors NATO ASI Series, Springer-Verlag, New York, pages 190-209. Available on-line at http://www.p-mmm.com/founders/UsingTestingBody.htm

Committee on Undergraduate Science Education. (1997). Science teaching reconsidered : a handbook. National Academy Press.

Duit, R. (2006). Students' and Teachers' Conceptions and Science Education: A Bibliography. Available on-line at http://www.ipn.uni-kiel.de/aktuell/stcse/stcse.html

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. Phys. Rev. ST Phys. Educ. Res. 2, 010105.

Dye, L. (2005). Students Use Clickers to Help Guide College Lectures. Available on-line at http://abcnews.go.com/Technology/DyeHard/story?id=727409&page=1.

Gregg, V.R., Winer, G.A., Cottrell, J.E., Hedman, K.E., & Fourneir, J.S. (2001). The persistence of a misconception about visions after educational interventions. Psychonomic Bulletin and Review **8**:622-626.

Harris, M. B. (1998). Basic Statistics for Behavioral Science Research. Needham Heights: Allyn & Bacon.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. Phys. Teach. **30**:141-158.

McKagan, S.B. & Wieman, C.E. (2005). Exploring Student Understanding of Energy through the Quantum Mechanics Conceptual Survey, PERC Proceedings 2005. CP-33. Available on line at http://cosmos.colorado.edu/phet/survey/QMCS/

Klymkowsky, M.W. (2005). Can non-majors courses lead to biological literacy? Do major's courses do any better? Cell Biol. Ed. **4**:196-198.

Light, P., Littleton, K., Bale, S., Joiner, R., & Messer, D. (2000). Gender and social comparison effects in computer-based problem solving, Learning and Instruction, **10**:483-496

McClymer, J.F. & Knoles, L.Z. 1992 Ersatz Learning, Inauthentic Testing. Journal of Excellence in College Teaching **3**:33-50. Available on line at http://bioliteracy.net/Readings/ErsatzLearning.pdf

Pelaez, N.J., Boyd, D.D., Rojas, J.B., & Hoover, M.A. (2005). Prevalence of blood circulation misconceptions among prospective elementary teachers. Adv Physiol Educ. **29**:172-81.

Pew Research Center Report (2005). Public Divided on Origins of Life. Available on line at http://pewforum.org/publications/surveys/religion-politics-05.pdf

Powell, K. (2003). Science education: Spare me the lecture. Nature **425**: 234-236

Redish,E. (2003). Teaching Physics with the Physics Suite, J. Wiley.

Seymour, E. & N.M. Hewitt (1997). Talking About Leaving: Why Undergraduates Leave the Sciences, Westview Press Boulder, CO.
Sokoloff, D.R. & Thornton, R.K. (1998). Assessing Student Learning of Newton's Laws: The Force and Motion Conceptual Evaluation of Active Learning Laboratory and Lecture Curricula, American Journal of Physics, **66**: 338-352.
Ward, D.L. (2003). The Classroom Performance System: The Overwhelming Research Results Supporting This Teacher Tool and Methodology. Available on line at http://www.einstruction.com/index.cfm?fuseaction=news.display&menu=news&content=showArticle&id=66.
Wood, W.B. (2004). Clickers: a teaching gimmick that works. Dev. Cell **7**: 796-798.

Supplemental Figure (not for publication): Chi Square analysis

Web Chi Square Calculator: Results

Raw numbers 35 males / 38 females

	conf. wrong	conf. right	total confident	Total
male	657	2944	3601	7202
female	782	3178	3960	7920
Total	1439	6122	7561	15122

Degrees of freedom: 2
Chi-square = 2.76310329331422
For significance at the .05 level, chi-square should be greater than or equal to 5.99.
The distribution is not significant.
p is less than or equal to 1.

Males values normalized to 38 (# of females)

	conf. wrong	conf. right	total confident	Total
male	713	3196	3909	7818
female	782	3178	3960	7920
Total	1495	6374	7869	15738

Degrees of freedom: 2
Chi-square = 2.90503136109425
For significance at the .05 level, chi-square should be greater than or equal to 5.99.
The distribution is not significant.
p is less than or equal to 1.