Ed’s Tools: A Web-based Software Toolset for Accelerated Concept Inventory Construction

Kathy Garvin-Doxas, Isidoros Doxas, & Michael W. Klymkowsky,
The Bioliteracy Project, University of Colorado, Boulder, CO 80303-0347

Published in "Proceedings of the National STEM Assessment Conference" 2006. D. Deeds & B. Callen, eds (2007). Pp. 130-139.

Abstract:

Ed’s Tools is a suite of web-based programs designed to facilitate the research that underlies, and the construction of concept inventory (CI) questions. Ed's Tools was developed for the Biology Concept Inventory (BCI) project, but has since been used in different fields, from engineering to space physics. It facilitates the collection, coding, and analysis of text based student data, as well as the synthesis of the results. Multiple, geographically separated coders can analyze and share data, and the results are available in a simple database of coded student responses. Ed's Tools is freely available via the BCI website, bioliteracy.net.

Introduction

For the last fifteen years or so, a revolutionary transformation has been taking place in the way Physics is taught at the college level. Physics courses at all levels, from the standard introductory course for non-science majors to upper level courses for physics majors, are being transformed from primarily lecture-based to a more "interactive engagement" style of instruction. At some schools this has led to the complete redesigned the classroom layout, while at others the emphasis has been more on the introduction of student response systems ("clickers") that facilitate real time interactions between instructor and students. Finally, a number of integrated instructional packages have been published (e.g. Tutorials in Introductory Physics, Prentice Hall; Workshop Physics, Wiley) that combine meticulously researched materials and tightly scripted instructional methodologies.

The revolutionary transformation of the way Physics is being taught can be traced in large measure to a particular catalytic event: the development of the Force Concept Inventory (FCI; Hestenes, et al., 1992). Concept inventories (CIs) are validated and reliable multiple choice instruments that explore students’ conceptual understanding in a give subject area; they provide researchers with a common instrument that can be used to measure the improvement in student achievement after a single course. More general tests, like the SAT or the GRE, have long been used as measures of achievement, but they are inappropriate for evaluating the change in student conceptual knowledge after a single course, and in general are not designed to specifically measure student mastery of particular ideas. Armed with pre- and post-instruction FCI data, Hake (1998) found that courses that use collaborative learning techniques (referred to generically as "interactive engagement") have learning gains that are often twice those seen in classes that use only traditional lectures. This study is commonly cited as the impetus driving the subsequent transformation of a large (and ever increasing) number of physics courses.

Spurred on by the impact of the FCI, a large number of efforts have begun to create similar instruments for other subject areas. A non-exhaustive list includes the Astronomy Diagnostic Test (Hufnagel, et al., 2000), the Thermodynamics Concept Inventory (Midkiff et al., 2001), the Natural Selection Concept Inventory (Anderson et al., 2002), the Fluid Mechanics Concept Inventory (Martin et al., 2003), the Geoscience Concept Test (Libarkin et al., 2003), the Basic Biology Concept Inventory (Klymkowsky et al., 2003; Garvin-Doxas & Klymkowsky, ms. in preparation), the Materials Concept Inventory (Krause et al., 2004), and the Chemistry Concept Inventory (Jenkins et al., 2004). At the same time, it is worth noting that the impact of these various concept inventories on their respective target fields has been rather less dramatic than that of the FCI. This raises a practical question, what factors determine whether a concept inventory has a significant impact on teaching and learning within its target area. Although a thorough discussion of this question is beyond our scope here, one can speculate as to the factors that lead to the widespread effects of the FCI; these include the championing of FCI data by prominent educators (e.g. Eric Mazur and Richard Hake) as a valid reflection of learning. In addition, one might argue that physics, as a discipline, is distinct in the near universal acceptance of key ideas (e.g. Newton's laws of motion and Maxwell's equations) as organizational principals in the introductory course sequence. In others areas, such as biology, finding a similar consensus upon which to focus conceptual assessment appears to be much more problematic; in our own experience, much time and effort is required simply to identify the foundational concepts to be evaluated.

That said, even with clear conceptual targets, CI research, construction, validation, and reliability requires considerable time and effort. The creation of the FCI (Hesteness et al, 1992) took almost a decade, not withstanding the fact that it addresses an extremely limited and well-accepted cluster of concepts in classical physics, namely Newton's laws of motion. Moreover, no substantial advance has been made in the amount of time, effort, and expense required to develop a validated, reliable, relevant, and widely accepted instrument. Precisely because of the effort required to produce them, existing instruments are jealously guarded to protect their validity. The FCI, for instance, although freely available, makes a point of admonishing users to not include its laboriously researched questions in any for-grade exams lest students' very efficient "answer dissemination" networks make the answers widely available, and so significantly degrading the validity of a very expensive instrument. Even, so, as the content of the a particular instrument becomes widely known, there is a serious possibility that instructors will, often subconsciously, begin to "teach to the test."

Concept inventories versus standard tests: While an experienced teacher may be able to assess a student's understanding of a particular concept through a relatively brief one-on-one examination, this approach is not feasible for the assessment of large number of students. In the same way, essay-type exams are equally time consuming to development and to analyze objectively, and it is often the case that such tests are "inauthentic" in the sense outlined by McClymer & Knowles (1992). The same reasoning that underlies the use of double-blind, placebo-controlled trials in the biomedical sciences also applies to educational assessment;

instructors have a vested interest, both personally and professionally, with regards to the evaluation of student learning – they cannot be assumed to be objective, no matter how hard they try.

Developing and validating reliable assessment instruments is an arduous and time-consuming task, qualitatively different from that involved in the generation of a standard exam. In the case of a multiple-choice type instruments, there are two main obstacles involved in designing a valid, reliable instrument: the first is the design of the questions and the second are the choices supplied to students. For example, the question cannot be so detailed and jargon-laden that it artificially limits responses, whereas the correct choice cannot provide irrelevant "clues" that flag it as correct. Similarly, the incorrect choices, known as distracters, should reflect commonly held student misconceptions – such distracters lure students away from the correct answer, whereas distracters that are obviously wrong degrade the instrument's validity, since they improve the odds of guessing the correct response, a phenomenon known as "construct irrelevant easiness".

Once questions and choices have been generated, it is critical that the instrument be validated, so as to avoid false readings (Messick, 1989). This involves extensive interviews, to ensure that students who score highly actually understand the concepts embedded in the test, while those who do not (and picked the distracters) actually hold the misconception represented by the incorrect answers. By repeated trial-and-error, during which the exact wording of each question and answer is finely tuned, an experienced team of content and assessment experts can increase the validity of a test score inference. Reliability has to do with the instrument's ability to consistently produce the same result when administered to the same population (independent, internal consistency measures like Cronbach's alpha, are used to measure the reliability of instruments) and can be degraded when the distracters are arbitrary, thus introducing random elements into subjects' responses. As with validity, an experienced team of content and instrument experts can produce a reliable instrument after only a few iterations.

Concept Inventory Construction and Ed's Tools: Developing questions, answers and distracters for a concept inventory has to be repeatedly anchored in students' conceptual understanding of the subject (Figure 1). It follows, rather rigidly, a pathway that ensures the adherence of the development process to what the students think, as opposed to what the developers (or content experts) believe they think. These steps consist of:

1) Assign essay questions to students. These are typically board questions designed to elicit open-ended responses that represent the spectrum of conceptual understanding of the responders on the subject matter. These responses are the start of a process of mapping the students' conceptual landscape, which is often poorly appreciated by instructors and experts in the subject matter.

2) Analyze language of respondents. The essays are coded (not "graded") for all concepts, both correct and incorrect, present in the answers. This "catalog of concepts" provides the raw material for questions, answers and distracters.

3A) Use student language to develop candidate questions and answer choices. This intermediate step leads to the use of largely verbatim quotes respondents' language as the possible answers in the development versions of the inventory. It accelerates the production time by allowing the developers to quickly isolate the type of language that elicits valid responses from the respondents.

3B) Produce written explanations for the purpose behind each question, and the concept each answer is meant to probe. This is a critical step, which requires coders to make explicit the connection between the question or answer they propose and the coded concepts.

4) Conduct validity tests. Validity testing involves a variety of standard methods, like interviews, both individual and group, think-alouds, etc. This process enables the interviewer to gauge how well the question and answer represents the student's conceptual understanding.

5) Refine questions and answers. The objective of this step is to produce a valid list of prevalent concepts, not to score correct answers.

6) Conduct reliability testing. This step consists of administering the candidate questions to a sample population and using standard statistical techniques to measure reliability.

7) Repeat the process (steps 3-6) until some criterion for reliability convergence is reached.

It was to facilitate steps 1 and 2 of this process that Ed's Tools was developed. In our first pass through a new concept area, we generate what we believe are broad, open-ended essay type questions that we expect will generate responses of between 100-200 words. Our goal is elicit responses that go beyond the usual rhetoric of science employed by our students. Often these questions fail in this regard, and must be reworded (see below).

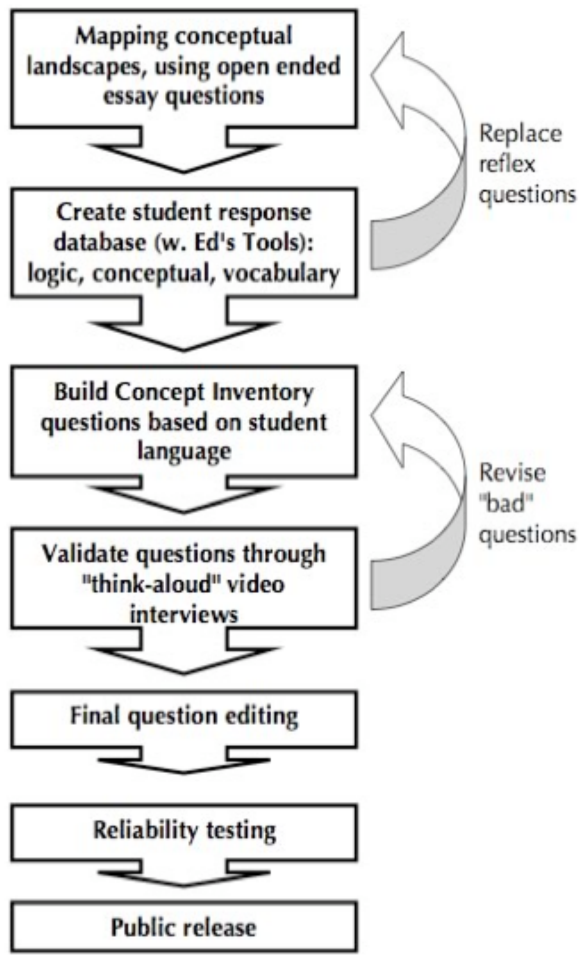

Figure 1: The Development Process

By its very nature, steps 1 and 2 require that we collect and analyze large numbers (thousands) of essays. Initially, we planned to use the usual means of data collection and analysis: a simple web program to collect data that could then be "dumped" into a computer assisted text analysis program (e.g., Nudist), and a standard statistical analysis program (e.g., SPSS) for tracking pilot test results. While this initially seemed reasonable, in practice it proved awkward to implement. More than half the functions in the text analysis programs were not useful for concept inventory development, and the functions we did need proved cumbersome, so much so that our initial round of coding was actually conducted using pencil and paper, followed by discussion and comparison of results, deciding on common coding categories, and then following the process again (Fig. 1). This was simply too inefficient a process for long-term use, since it required too much face-to-face time. We quickly realized that we needed tools that better met our needs and increased our efficiency; preferably tools that were platform-independent; accessible to all participating coders (faculty volunteers and graduate students working with us); web-based, so that results could be combined and compared easily; and easy to learn and use for non-social scientists and inexperienced graduate students. We therefore developed Ed's Tools (see below), a program that meets all of these needs (and more) since it allows for essay data collection, iterative rounds of coding, actual grading, and tracking student participation by course instructors when they elected to give students credit for responding to essays (The program was named after the undergraduate who coded the program). Funding for Ed's Tools was provided by the NSF as part of the BCI project. The result is a web-based tool currently being used by STEM educators in the biological sciences, astronomy, and space physics who have similar needs and are interested in student conceptual understanding. It is hosted by the Center for Integrated Plasma Studies (affiliated with the Department of Physics at the UC Boulder) and available at: https://solarsystem.colorado.edu/conceptInventories/ or through the bioliteracy web site (http://bioliteracy.net).

Student response analysis: We analyzed essay responses using content analysis and beginning without *a priori* assumptions. This was facilitated by the fact that non-instructors did most of the coding. Content analysis is a technique designed to identify recurring patterns in key words, the relationships articulated or implied by the way words and phrases are consistently used, and oppositions between words and phrases (e.g., de Sola Pool, 1959; Holsti, 1969). The approach involves the use of multiple coders reviewing essay data and looking for themes and patterns. This gives us a picture of where students are in terms of their thinking and assumptions about specific concepts. Content analysis is reasonably simple to perform using Ed's Tools.

After multiple individual analyses by each coder, patterns and themes were discussed and sorted into clusters or groupings with similar meaning, leading to categories that accurately represent the conceptual landscape as it is experienced and understood by students. Recurring patterns both among individual questions and across different questions made it possible to identify those areas where students experience common difficulties in their understanding. For example, content analysis of essay data indicated that students are able to answer certain types of questions about natural selection correctly (questions asking them to define, label and categorize), but were unable to demonstrate conceptual level understanding of natural selection when asked to explain, recognize, or apply their understanding (Garvin-Doxas & Klymkowsky, in preparation)

The research questions addressed during the initial phases of essay analysis are: 1) What recurring themes, patterns, and language do students employ when discussing these broad topics; 2) Are there patterns in their responses that span different questions; 3) What do the recurring themes, patterns, and language tell us about student understanding in these areas; 4) Do any of these recurring themes, patterns, and/or language indicate common misconceptions or do student responses demonstrate that the majority have solid conceptual understanding in these areas? If results of initial essay analysis are meaningful and complete, we move on to the next step, student interviews, if not, new essay questions need to be developed and asked (step 1). Three types of interviews are involved, again in an iterative process. The initial type of interview is broad and thematic, rather than structured, and focuses on the validation of the interpretation of essay-based content analytic results. These interviews may be held in a small group in order to provide peer discussion from which to draw on, or they may be individual interviews, and deal with a restricted set of themes.

Armed with a reasonable certainty of student conceptual level understanding and strong indications of the areas where they commonly hold misconceptions as well as how they articulate them, we move on to the second type of interview. These are moderately scheduled think-alouds where students participate in question development and the validation of research results from essay coding and thematic interviews by talking through their thinking processes as they read potential questions and responses. Using as much student language as possible, we create a set of candidate questions along with student expressions of common misconceptions that they then talk through for us. Our research questions at this phase are: 1) Do the students understand the question in the way we intended; 2) when they read each distracter, do they interpret them in the way intended; 3) when they select a response, does it represent their conceptual understanding; and 4) do we need to modify questions and/or potential responses so that they more accurately reflect student conceptual understanding (particularly their misconceptions)? Again, this is an iterative process, and we may find at this point that we must develop new essay questions or conduct more thematic interviews.

Finally, we refine questions and potential responses based on all interview data, when possible, we compare to essay data (often, there is a large gap between essay responses and the product developed after the second type of interview), and we produce a cluster of questions aim at a particular conceptual target. This is followed by the final type of student interview, based on scheduled think-alouds. Students first fill out the instrument and then participate in the validation process by talking through what they understood each question asked and how they selected their responses. Our research questions at this phase are: 1) do students interpret the questions and potential responses in the ways intended; and 2) does the students' selected answer accurately represent their conceptual understanding both at the level of the individual question and at the level of the conceptual cluster?

The final step is the administration of the refined version of the instrument (FIG. 1). We look for quantitative results, patterns among answers (e.g., unnecessary questions), test-re-test reliability, etc. and then begin pilot testing. This type of validation processes relies heavily on a qualitative component where language and meaning are featured, and is meant to ensure that students believe questions and responses hold the same meaning that developers intended. Reliability is a statistical measure of the repeatability of results in like populations. We find that type 1 interviews typically require between 5-10 students; type 2 interviews benefit from between 10-15 students; and type 3 interviews require no more than 30 students. This is especially true when the students are offered a reasonable amount of money for their time (about 1 hour), so that a representative sample of students is examined.

Throughout cluster and question development, we ensure that the research justifications for each question and potential response are clearly articulated and met. Figure 2 shows an example (Question 25 from the BCI) of the resulting articulation of the conceptual understanding that it is intended to explore is required, and each potential response is directly traced to commonly held misconceptions supported by essay and interview research. The model is also designed so that the purpose of the question and distracters will be clear for all users (who can thus make informed decisions about including a question in their customized instrument).

Q25. Imagine that you are an ADP molecule inside a bacterial cell. Which best describes how you would manage to "find" an ATP synthase so that you could become an ATP molecule? *This question is designed to test whether students understand that diffusion is caused by random motion of molecules. Our research shows that most students do not understand that molecules are in constant motion. Although they understand the concept of gradients and moving down a concentration gradient, they don't understand what causes this movement.*

a. I would follow the hydrogen ion flow.
Students who choose this answer think that ADP somehow can identify where a hydrogen ion gradient is. In our research, students never explain how an ADP would sense an H+ gradient. Many students use language that suggests they think ADP is able to actively seek out an ATP synthase. For example, ADP is described as "looking for" or "noticing." Although the answer is incorrect, students who select may understand correctly that there is a hydrogen ion gradient. The may also correctly understand that the gradient is causing hydrogen ions to flow through the synthase and across the plasma membrane to the area where hydrogen ions are less concentrated.

b. The ATP synthase would grab me.
Students who select this answer think that an ATP synthase senses the presence of ADP and actively grabs it. Excluding some students who said ATP synthase has a receptor that recognizes ADP, most students do not explain how the ATP synthase does this grabbing. One again, students who select this answer may believe that molecules have the ability to actively seek out or choose other molecules.

c. My electronegativity would attract me to the ATP synthase.
Students who select this answer think that charges cause the ADP and ATP synthase to be attracted to each other. Based on what students know about molecular interactions, it's a good guess. However, it seems to be just that – a guess – because students don't say what charge the ATP synthase is or that the ATP synthase is oppositely charged.

d. I would be actively pumped to the right area.
Students who select this answer think that the ADP is somehow placed in the correct area so that it is close to the ATP synthase. Students may have no explanation for how this occurs.

e. Random movements would bring me to the ATP synthase.
This is the correct answer. In other words, ADP finds ATP synthase by the random motion of ADP molecules.

Figure 2: Question Development Model

The structure of Ed's Tools: Ed's Tools provides a facile system for the collection and analysis of **<u>any</u>** kind of student text on-line (cf. Figure 3), or any data that can be translated into text (i.e., interview transcripts) and the system allows responses to be transferred into SPSS for analysis. In future versions, we plan to include the ability to upload graphics files since these can also reveal important student assumptions.

Ed's Tools was designed and built on a "just-in-time" basis, meeting project needs as they arose, and so there are still areas that would benefit from additional refinement: in its current form, it contains a posting-place for assigned or optional essay questions; a coding site that allows for multiple passes; provides a means for comparing coding results from more than one coder; and a site where a selected series of responses can be viewed (e.g., all responses for a particular question can be selected or a sample of responses can be seen and printed if needed). We have found that students, instructors, and researchers can easily master Ed's Tools, and Ed's Tools makes it fast and easy to collect student essay responses and facilitates the iterative process of coding those essays (interview data is handled in the same way when transcribed). Our ability to rapidly identify expanded BCI concept clusters and create CI questions is largely dependent upon Ed's Tools.

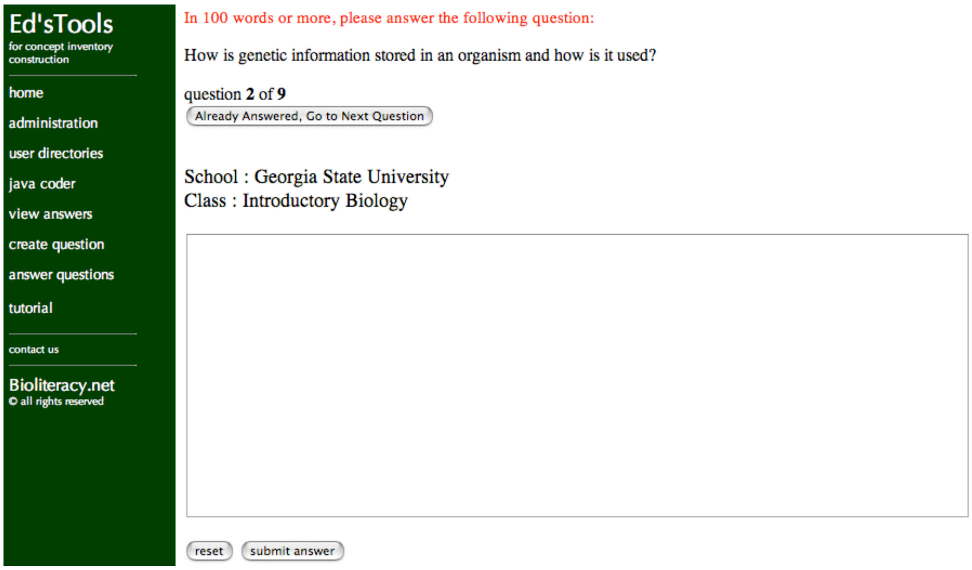

Figure 3: The input field. Instructors can assign any combination of questions to any class. Answers are immediately imported into the answer database.

We have found that Ed's Tools can greatly facilitate collaboration during the iterative process of coding student expressions (verbal or written) of their understanding of the biological sciences. Figure 4 illustrates the Ed's Tools workspace for coding textual data. An individual coder logs on (more than one coder can be logged on at a time without interfering with one another) and then selects a question s/he wishes to work on. The student response then becomes visible within the coding window (font size can be changed for better legibility) and the question is displayed below. The researcher reads through the essay in an effort to identify recurring patterns within the response (usually we need to read through 5 or more essays before we begin to see patterns). While reading through each text, individual coders highlight the portions of text that represent a recurring pattern using the mouse, assign each one a color, and type in a label naming each category of response, theme, or pattern. Each category is identified in the right hand text box, together with its associated color. Multiple categories can be assigned to the same text. During first round coding, these labels tend to be fairly detailed (rather than broad, abstract ideas), so that nothing that may prove to be important is lost.

Once entered, each coder's labels remain with the question set as s/he moves along to new student responses. New labels can be added at any time during the process and text can be labeled as representing multiple categories, themes, and/or patterns. Second round coding generally remains an individual activity in which several of these detailed labels are subsumed under broader, more inclusive (and usually scholarly) names, and labels that do not recur get dropped. None of the first round coding is lost during this process. Third round coding generally involves looking at what the other coder(s) have done with the same data set; considering areas where they overlap and/or disagree; arriving at even more inclusive labels for recurring patterns, categories, and themes; and finally, applying these “third level” codes to an even larger portion (if not all) of the data set. Agreement among coders is much more quickly arrived at using Ed’s Tools because each can see what the other coders are doing and have done, consider it asynchronously, and then make group decisions either face-to-face or on-line. The color-coding scheme, with the ability to call up or print out student language representing each coding category alone, makes the entire process faster.

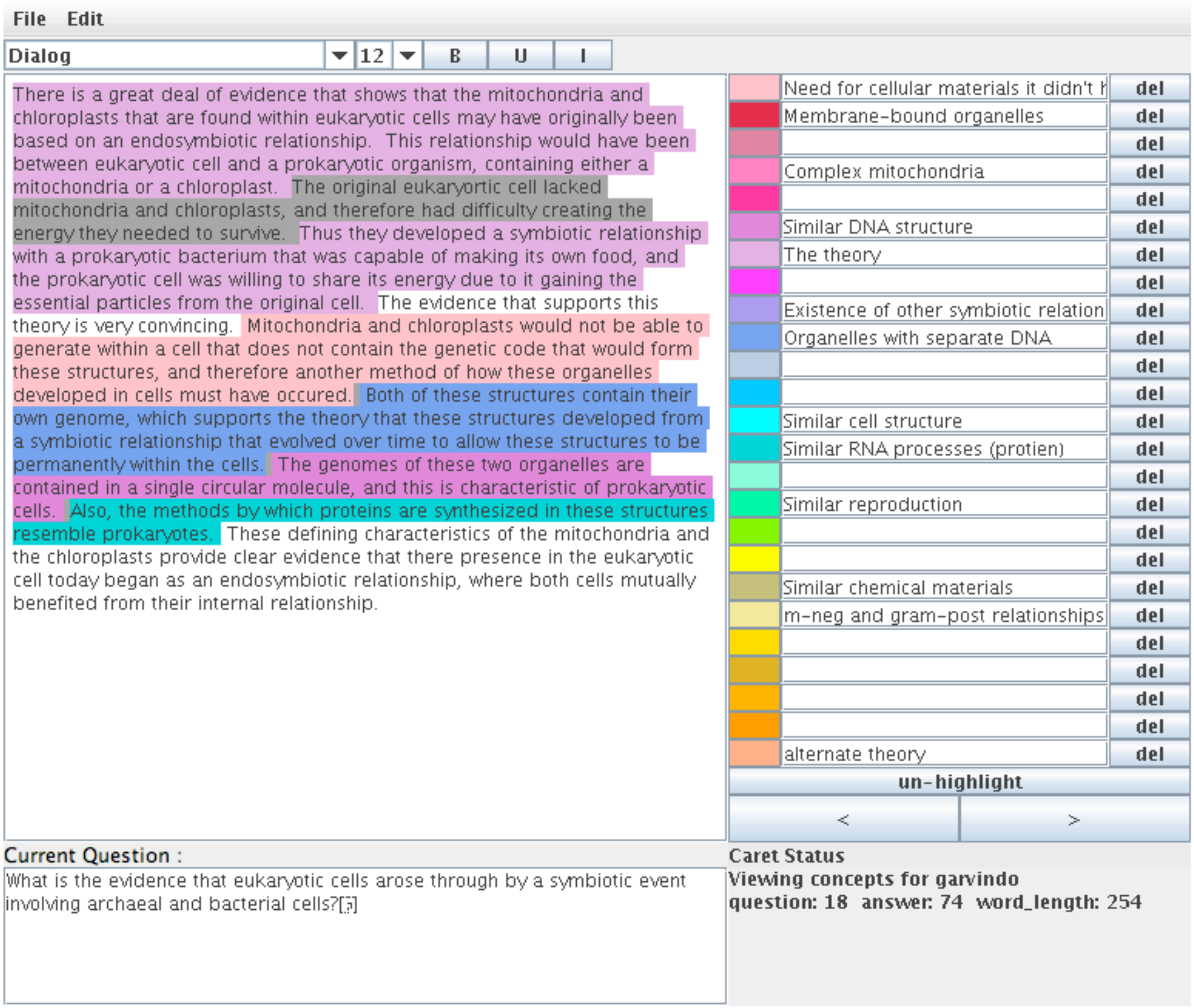

Figure 4: The Concept Coder. The text is imported from a database that associates a set of codes with a coder. Codes can be edited, viewed, or imported. Only the author can change a code, but others can view it and import it (i.e. adopt it for their own).

After the coding of a question is completed, the language (the responder’s phrases) that the coders have identified with a particular concept can be aggregated for all (or some) responses using the aggregation tool (Figure 5). The developer specifies the name of the coder, the question, and the concept (top), and the tool aggregates all the text that this particular coder associated with that concept (bottom). A number of permutations of these capabilities are possible, e.g. show the text as tagged by all coders, or show all the concepts this coder tagged for this question, etc. Once the language is thus aggregated, it can be used to assign candidate answers (verbatim) to possible Inventory questions. It is important that the answers be used more or less verbatim at this point, as it is the natural language of the students that best represents their misconceptions, and is therefore most likely to resonate with their actual thinking.

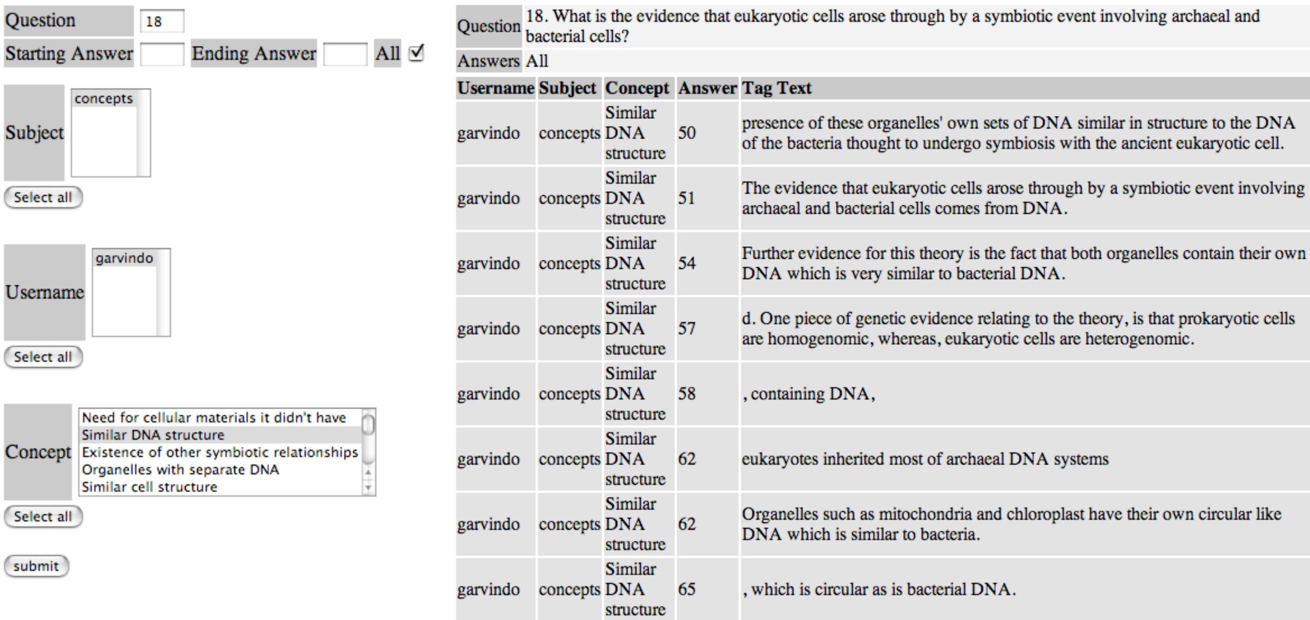

Figure 5: The language aggregator. Investigators can select from the database the student language that each coder associated with a given concept (top), which is aggregated (bottom), facilitating the construction of the first draft of an Inventory.

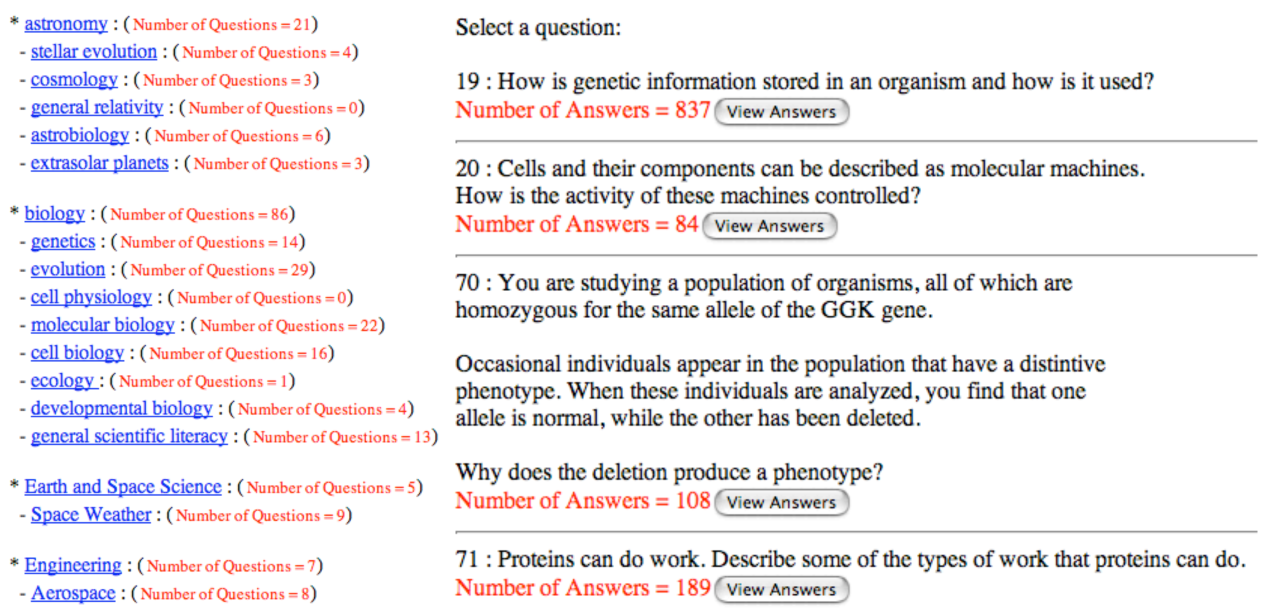

Figure 6: Approximately 10,000 essays, on subjects from Astronomy to Engineering, have been collected using Ed's Tools for 28 different classes. A sample 10 answers for each question can be perused at the BCI website.

Although Ed's Tools were developed for the BCI project, it is currently also used for the construction of instruments in Astronomy and Geoscience (cf. the Space Weather Concept Inventory, SPCI, at solarsystem.colorado.edu/conceptInventories/SPCI), as well as an essay submission and grading tool for courses in engineering. We have collected in all, over 10,000 essays. The topics covered, as well as the number of essays available for each question can be seen on the public website (cf. Figure 6). A sample of 10 essays for each question can also be perused at the same site.

Conclusions: Concept Inventories have catalyzed the transformation of teaching of physics, and hold the promise of doing the same for disciplines from biology to engineering, but Concept Inventory construction is a long and laborious process, mainly because the validation processes relies heavily on a qualitative component where language and meaning are featured. Ed's Tools greatly facilitates the initial validation stages of the process, and we invite others, both researchers and teachers, to use them.

Acknowledgements: This work is part of the NSF-funded Building a Basic Biology Concept Inventory project. While we have had a number of helpful participants, we are particularly grateful to Ed Svirsky for programming, Sara Pallas and Richard Cyr for their support in the early phase of the project, and the cooperation and support of faculty in Molecular, Cellular and Developmental Biology, and the Discipline-based education research group at the University of Colorado, Boulder.

Literature cited

Anderson, D.L., Fisher, K. M., and Norman, G. J., Development and Evaluation of the Conceptual Inventory of Natural Selection, J. Res. Sci. Teaching, 30: 952-978 (2002).

Hake, R., Interactive engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *American Journal of Physics, 66*, 64 (1998).

Hestenes, D.,Wells, M., and Swackhammer, G. Force Concept Inventory, *The Physics Teacher, 30*, 141-158, (1992).

Hufnagel, B., Slater, T., Deming, G., J. Adams, Adrian, R.L., Brick, C., and M. Zeilik, M. Pre-Course Results from the Astronomy Diagnostic Test, Publications of the Astronomical Society of Australia, 17(2), 152 (2000).

Jenkins, B.E., Birk, J.P., Bauer, R.C., Krause, S., and Pavelich, M.J. "Development and Application of a Chemistry Concept Inventory", Symposium on Research in Chemistry Education," 227th National Meeting of the American Chemical Society, Anaheim, CA, Mar. 28-Apr. 1 (2004).

Klymkowsky, M., Garvin-Doxas, K., and Zeilik., M., Bioliteracy and Teaching Efficacy: What Biologists can Learn from Physicists, Cell Biology Education, 2, 155-161 (2003).

Krause, S., Tasooji, A., and Griffin, R. "Origins of Misconceptions in a Materials Concept Inventory From Student Focus Groups," Proceedings, ASEE Annual Conference (2004).

Libarkin, J., and Anderson, S.W., The Geoscience Concept Test: Linking grounded theory, scale development, and item response theory; EOS, Transactions of the American Geophysical Union, v. 84, Abstract ED22E-06 (2003).

Martin, J.K., Mitchell, J., and Newell, T., Development of a Concept Inventory for Fluid Mechanics, Proceedings, Frontiers in Education Conference, Boulder, CO, USA (2003).

McClymer, J.K. & Knoles, L.Z. 1992. Ersatz Learning, Inauthentic Testing. Journal of Excellence in College Teaching 3:33-50.

Messick, S., Validity. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan (1989).

Midkiff, K.C., Litzinger, T.A., and Evans, D.L., "Development of Engineering Thermodynamics Concept Inventory Instruments," Proceedings, Frontiers in Education Conference, Reno, Nevada, 10–13 October (2001).

Tutorials in Introductory Physics, Prentice Hall, Upper Saddle River, NJ (2002).

Workshop Physics Activity Guide, Mechanics I: Kinematics and Newtonian Dynamics, 2nd Edition, Wiley (2004).